MULTIVARIATE NUMERICAL ANALYSES OF *RAOULIA* SUBGENUS *RAOULIA*

A THESIS

SUBMITTED IN PARTIAL FULFILMENT

OF THE REQUIREMENTS FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF PLANT AND MICROBIAL SCIENCES

UNIVERSITY OF CANTERBURY

C.M. FRAMPTON

UNIVERSITY OF CANTERBURY

1988

ABSTRACT

Through the sequential analysis of a data set, comprising 86 OTUs of *Raoulia* subg. *raoulia* measured on 98 characters, different multivariate numerical manipulations are performed, compared and assessed.

The creation of additional characters derived from sampled characters (ratios) is investigated. The univariate distributional properties of these derived characters, are seen to be mainly non-normal, although strategies for minimising this in terms of the choice of numerator and denominator are advocated. To include a ratio and its two constituent characters in a data set will invariably lead to the multiple inclusion of the same information beyond a level that might occur with normal inter-character correlations in a biological data set. Through the exploration of the multiple and bivariate correlation coefficients it is shown that simple calculations will reveal an optimum strategy in terms of removing one of the three characters. Using three shape ratios independently as examples and developing a discriminant function for the 13 *a priori* defined species in the data set, based on the three characters numerator, denominator, and the ratio, it is shown that the ratio *per se* has the superior correlation with the grouping variable in each instance.

Using the multivariate moments of skewness and kurtosis, the multivariate normality of the taxon based on the 63 continuous characters was assessed. This normality of the taxon was seen to be dependent primarily on the character number and the status of the OTUs most distant from the centroid. The repercussions of the removal of the outlying OTUs and the status of these six outliers was further explored by cluster analysis. Using a reduced 39-character data set the individual group normality of both portions of three successive dichotomies was assessed. It was seen from this that a reasonable range of Mahalanobis D^2s about a centroid was essential for multivariate normality. It was further shown, via a pair-wise discriminant analysis on each of the three partitions, that OTUs disturbing the multivariate normality of a single group are not necessarily those that are misclassified within the prescribed groups.

The inter-group associations revealed by the canonical variate plots and the jackknife Mahalanobis D^2s indicated the possibility of amalgamating a number of the species groups. The detection of outlying OTUs on the basis of large relative minimum jackknife Mahalanobis D^2s was compared with the detection via the earlier single-group analyses this showed that even apparently

extreme outliers were providing some not inconsiderable stability to their hypothesised groups, and that their removal could be an extreme course of action.

In order to reassess the above changes using the entire data set and to reach a conclusive grouping strategy a new method was proposed as being appropriate to such circumstances . This method allows for the independent summation of the univariate χ^2 and F ratios approximated by χ^2 values, of each character based on a given grouping strategy. These values were recomputed for an alternative strategy and the difference between the sums compared to the χ^2 distribution, with the change in the degrees of freedom as the degrees of freedom.

Given the 'final' groups defined by the previous analysis, characters were extracted to form a diagnostic hierarchy of dichotomous subdivisions.

CONTENTS

PAGE

CHAPTER ONE: INTRODUCTION

INTRODUCTION TO NUMERICAL TAXONOMY 1

NUMERICAL PROCEDURE 4

SELECTION OF CHARACTERS AND OPERATIONAL TAXONOMIC UNITS 7

CHAPTER TWO: RATIOS

INTRODUCTION 10

METHODS 11

RESULTS 13

DISCUSSION 17

CONCLUSIONS 20

CHAPTER THREE: MULTIVARIATE NORMALITY

INTRODUCTION 22

METHODS 24

RESULTS 26

DISCUSSION 36

CONCLUSIONS 43

CHAPTER FOUR: DISCRIMINANT ANALYSIS

INTRODUCTION 45

METHODS 48

RESULTS 49

DISCUSSION 61

CONCLUSIONS 65

CHAPTER FIVE: COMPARATIVE ASSESSMENT OF GROUPING STRATEGIES USING ALL CHARACTERS

INTRODUCTION 66

METHODS 68

RESULTS 72

DISCUSSION 72

CONCLUSIONS 75

CHAPTER SIX: DIAGNOSTIC CHARACTERS

INTRODUCTION 77

METHODS 78

RESULTS 79

DISCUSSION 83

CONCLUSIONS 84

CHAPTER SEVEN: CONCLUSIONS 85

ACKNOWLEDGEMENTS 89

REFERENCES 90

APPENDICES

A: OPERATIONAL TAXONOMIC UNITS: SPECIES NAMES 98

B: OPERATIONAL TAXONOMIC UNITS 99

C: CHARACTER SET 101

D: ORIGINAL COMPUTER PROGRAMS 108

LIST OF TABLES

TABLE		PAGE
1.	The explained variation in a ratio (Z:X/Y) as a result of X and Y ($R^2_{z.xy}$) for different variable correlation structures.	14
2.	Parameters for three ratios and their constituents.	15
3.	Parameters for shape ratios and their reciprocals.	16
4.	Mahalanobis distances of all OTUs to the centroid of the subgenus.	28
5.	Mahalanobis distances for the '*hookeri*' and 'non-*hookeri*' groups.	33
6.	Mahalanobis distances for the 'diploid' and 'non-diploid' groups.	35
7.	Mahalanobis distances for the 'riverbed' and 'non-riverbed' groups.	38
8.	Mahalanobis distances for the initial groups.	51
9.	Mahalanobis distances for the initial groups with deletions.	53
10.	Mahalanobis distances for the initial groups with amalgamations.	58
11.	Mahalanobis distances for the initial groups with amalgamations and deletions.	59
12.	Changes in chi-square and z-scores associated with different grouping strategies.	73
13.	Diagnostic characters for the final species groups.	80
14.	Abnormal characters for outlying OTUs.	82

LIST OF FIGURES

FIGURE		PAGE
1.	Dendrogram for all OTUs using continuous characters.	29
2.	Dendrogram for 80 OTUs using continuous characters.	31
3.	Plot of canonical variable 1 for *'hookeri'* and 'non-*hookeri*' groups.	34
4.	Plot of canonical variable 1 for 'diploid' and 'non-diploid' groups.	37
5.	Plot of canonical variable 1 for 'riverbed' and 'non-riverbed' groups.	39
6.	Plot of canonical variables 1 and 2 for original 13 species groups.	50
7.	Plot of canonical variables 1 and 2 for 12 species groups with 5 OTU deletions.	55
8.	Plot of canonical variables 1 and 2 for 9 species groups.	57
9.	Plot of canonical variables 1 and 2 for 9 species groups with 5 OTU deletions.	60
10.	Plot of canonical variables 1 and 2 for final 12 species groups.	81

CHAPTER 1

INTRODUCTION

"Where shall I begin please your majesty?' he asked.
'Begin at the beginning,' the king said, gravely, 'and
go on till you come to the end: then stop.'"
LEWIS CARROLL

The publication of *Numerical Taxonomy* (Sokal & Sneath, 1963) heralded the beginning of an era for the application of numerical methods as an aid to solving problems in systematics. The book itself did not present previously unpublished statistical techniques, but gave credence to the possibility of the application of such techniques by systematists in a number of disciplines.

The term 'numerical taxonomy' was coined and defined as " ...the grouping by numerical methods of taxonomic units into taxa on the basis of their character states." (Sokal & Sneath, 1963). 'Numerical' implies the translation of character forms into a numeric state, thus making them available for mathematical manipulation. The term 'taxonomy', defined by Simpson (1961) as "...the theoretical study of classification, including its bases, principles, procedures, and rules", is used rather than the more general term 'systematics', which Simpson (1961) defined as "...the scientific study of the kinds and diversity of organisms and of any and all relationships among them." The individual units to be characterised in the context of numerical taxonomy, because they may be of no particular taxonomic rank but are assumed to be of a constant rank within a particular study, were defined as 'Operational Taxonomic Units' (OTUs), that is, "...the lowest ranking taxa employed in a given study." Systematic fields exist that are exclusive of numerical taxonomy but these can employ similar statistical methodology. Potentially there are a great number of these, wherein the categorical grouping criterion is not taxonomic but rather derived from a specific systematic criterion, but which still lead to a 'classification' that can be tested and characterised. The development and application of more statistical techniques today makes those of numerical taxonomy a large subset of those pertaining to numerical systematics.

The relationships revealed by numerical taxonomy are traditionally phenetic. That is, a measure of 'overall phenotypic similarity' is derived, given the states for a selection of characters from two specimens. No empirical assumptions regarding the phylogenetic relationships can be made if no deliberate attempt is made to distinguish characters of genealogical value, i.e. uniquely derived character states. This is not to say that phylogenetic inferences cannot be made from phenetic results based on overall similarity, but given that all measurements are made on extant specimens, the methods deriving overall similarity do no more than imply the likelihood of certain evolutionary pathways and do not distinguish similarities due to parallelism and reversion. They cannot lead to definitive conclusions on such pathways. Thus within the framework of overall similarity one is specifically dealing with phenetics not phylogenetics.

The nature of taxonomy generally, dealing as it does with the definition of biological populations, intuitively relates to the population as the fundamental basis of parametric statistics. Thus the individual and combined characters of any numerical study are linked with the biological nature of populations. Indeed, the composite form of correlated characters characterising a particular taxon, the fundamentally conservative nature of this form and its subsequent geometrically determined relationship with the environment are integral to the evolutionary process, orthogeny. This process itself leads to the creation of taxa that can be clearly delimited on the basis of their shared fundamental forms eg. Compositae, Gramineae. Given these natural associations, the notions of objectivity and repeatability traditionally associated with statistics could be incorporated into taxonomic studies. Early claims of the general significance of these qualities (Dupraw, 1965) have now been tempered by findings from many studies in a number of different disciplines which indicate that problems continue to arise and always will occur in the process of representing a phenotype in a form capable of being manipulated by a computer.

As more techniques have become available, ascertaining the appropriateness of any specific one to an individual study has become an increasingly difficult problem. Attempts to overcome this have led to a situation in which the success of any numerical study is likely to be assessed on the basis of rather arbitrary definitions of the 'best' biological solution. This tendency has been at the expense of the biological 'objectivity' and 'repeatability', and has highlighted the lack of a successful interface between the empirical (biological science) and the theoretical (statistical significance).

The above problems aside, it is clear that in the processes of pattern recognition in the relationships between delimited groups of a specific pattern and in the identification of the features which characterise a group of a specific pattern, the computer and the statistical techniques it makes available have an obvious and important role to play. This general point was clear when computers were the size of small buildings ("...statistics is today so intimately related to most aspects of systematic endeavour that it should be a required course for students training to be Systematists..." (Sokal, 1965)) and its pertinence is even more pronounced today.

GENERAL NUMERICAL PROCEDURE

"For he, by geometric scale,
Could take the size of pots of ale;...
And wisely tell what hour o' th' day
The clock doth strike, by algebra."
SAMUEL BUTLER

The usual procedure adopted by numerical taxonomists can be broken into specific stages, each with its largely independent array of appropriate techniques.

1]The characters and operational units are chosen and all combinations recorded. The non-statistical problems associated with this involve the delimiting and adequate sampling of the O.T.U. group, the choice of characters and *a priori* character weightings. These problems are not addressed in this thesis, but they are well described by Sokal (1965) and Sneath & Sokal (1973). Problems occurring at this stage that can be investigated statistically include sample size, character variability, construction of ratios as additional characters, logical and empirical associations of variables and the coding to a linear form of ordered discrete characters.

2]A measure of similarity or dissimilarity is calculated. A general overview of the available techniques is given by Sneath & Sokal (1973). Conventionally, despite the array of methods available, this step is not one of major concern. The various methods are all dependent on the type of data and the choices beyond this can normally made on sound empirical grounds.

3]Given the measures of either character or O.T.U. association, a representation of the p-dimensional structure must be formed in order to assess the O.T.U. relationships. Broadly, the techniques involved here are of either sequential clustering or non-sequential clustering (ordination). The selection of techniques within these two groups is in no way prescribed. Many are data-type dependent, but beyond this, arbitrary decisions are usually made by workers on the basis of previous efforts in their field of study. Many of the techniques have been shown to have empirical weaknesses in certain facets, for example exaggeration of large inter-O.T.U. distances by ordination and overemphasis on small inter-O.T.U. distances by cluster analysis. The overall impact of such weaknesses, however, is not clear until some picture of the OTU structure is

formed. Many of the techniques are statistically related, and some indication of the robustness of hypothesised associations is given by their repeatability with different techniques. A comprehensive coverage of all aspects of sequential cluster analysis is given by Anderberg (1973) and a good exposition of q-mode ordination can be found in Williams (1976). All forms of cluster analysis, both sequential and non-sequential, are hypothesis generating.

4]An assessment of the OTU relationships as represented is made, and taxa are delimited. In practice this step tends to be the least objective. The major problem is that in many data sets there is no one definitive solution, no single notion of 'best fit'. In the past phenon lines were drawn, in part utilising the biological edict that all pairwise differences between taxa of the same rank should be approximately equal. More realistic thinking has subsequently made it clear that such attempts at objective classification are not appropriate. Many data sets are hierarchical in terms of O.T.U. structure, and thus stopping rules and multivariate analysis of variance should be looking for clearly defined local maxima, in terms of optimising the between-within variance ratio. Under this criterion higher hierarchical disjunctures will dominate calculations. In reality, the 'success' at this stage should be a reflection of the faith the investigator has in his data set as a true representation of the phenotype, given that there is statistical conclusiveness. The tendency, however, is to take parts of the results as supporting evidence for *a priori* hypotheses and to neglect the rest. A more scientific approach is required whereby potentially supporting evidence that does not eventuate must be looked upon as contrary evidence. Statistical conclusiveness, itself requisite for biological conclusiveness, can be gauged by eigen values for non-sequential clustering and by cophenetic correlations (Sokal & Rohlf, 1962) or by techniques such as that of Hartigan (1967) for sequential clusters. These methods can then be followed by the hypothesis testing techniques such as MANOVA.

5]Character states unique to generated sub-groups are isolated for descriptive and diagnostic purposes. In practice, when there is no contention concerning the subgroup composition, this is a comparatively straightforward statistical procedure. Nominal discrete-state characters can be placed into contingency tables and analysed by the Chi-square and other related statistics. Numeric continuous and discrete state variables can be analysed by simple one-way ANOVA to find the variables maximising inter-group diagnoses. If these results are not immediately rewarding, then one can reconstruct the grouping variable into several dichotomous hierarchical variables, in the

manner of a diagnostic key. Then pair comparisons can be made at each level in the hierarchy and characters giving the best separation at each level extracted. If these methods fail, the practically less useful technique of stepwise discriminant analysis, either pairwise or multiple-group, can be used to derive character combinations that uniquely distinguish subgroups. This technique is most useful when there are latent factors, rather than the characters employed in the analysis, that are diagnostic, e.g., leaf size or the form of a floral structure. Many numerical studies come immediately to this step, asking the question "Can a data set of this form support my *a priori* classification ?". In this instance stepwise discriminant analysis is particularly appropriate, provided that one is aware of the inherent bias to regenerate the hypothesised groups. An elaborate technical exposition of discriminant analysis and related fields is provided by Lachenbruch (1975).

In following the above procedure, any numerical taxonomist will inevitably be confronted with dilemmas and decisions. These problems may be divided into two groups:

1]The adequate representation of the biological field of study in a numerical form. Included in this are the problems of phenotype and population sampling.

2]The appropriateness of specific statistical techniques to the problem. This field encompasses both the biological implications of statistical assumptions and the practical applicability of likely statistical conclusions.

Obviously these two areas are related in that the degree of phenotype and population sampling will influence the usefulness of results, while population sampling will have a direct influence on the normality and homoscedasticity of the samples. In effect, both areas require biological and statistical input. In the scientific studies of today both of these problems can be easily ignored, in that 'canned' computer packages will analyse data of any form and yield statistical conclusions. The usefulness of these conclusions is a direct consequence of attention to the two areas outlined. It is through the systematic numerical analysis of a taxonomic problem, incorporating the procedure outlined above, that this thesis aims to address a number of the frequently encountered but often ignored problems of numerical taxonomy. It is anticipated that the results of this study will have implications beyond the field of numerical taxonomy.

THE SELECTION AND ACQUISITION OF CHARACTERS AND OTUS

"It is a capital mistake to theorize before one has data." ARTHUR CONAN DOYLE

(i)OTU selection

The selection of OTUs for a numerical study is dependent primarily on the aims of the study. The major restriction is that each OTU should be approximately the same in terms of its hypothesised position within a taxonomic hierarchy. Thus, if one is examining the inter-specific relationships within a genus, it is invalid to include a genus from another taxon as an OTU. The reason for this is that the condition of character stability within each OTU is rarely met therefore, character averages and character 'representatives' are frequently used. If the OTUs are of approximately the same taxonomic rank, then one may generally assume that the amount of intra-OTU variability will be more equivalent than in OTUs of a different rank. Therefore, any classifications that result from the study will be such that a constant level of variability is excluded from each delimited taxa. If this level of variability is not constant, one can only approximate the OTU relationships, having no definitive measure of group overlap. The only other condition on OTU selection is that within the defined aims of a particular study all appropriate OTUs should be included. If one is exploring the intra-specific relationships within a genus, then all species should be collected. To merely collect the readily available species is likely to lead to spurious associations as a result of the lack of influence of the non-collected OTUs, which are required to complete the picture.

The data set to be analysed in this thesis as an example of the numerical methodology comprises 86 OTU's of the group *Raoulia* subg. *Raoulia*. All 14 species as defined by Ward (1982) are represented with varying sample numbers. The individual OTUs each represent a single specimen (Appendix A, Appendix B).

(ii)Character selection

It is not the purpose of this thesis to enter into the debates concerning requisite character number or character definition. There are, however, other problems associated with the successful sampling of a phenotype. The assumptions made with regard to character selection are dependent

to a large degree on the form of the study. In this regard, numerical taxonomy has a specific set of assumptions.

(1)The phenotype of an OTU can be parameterised so as to obtain a measure of its overall form. Therefore, to measure overall form, although morphological characters usually predominate, biochemical, physiological, and ethological characters are equally admissible. A character itself, or more precisely a unit character (Sneath & Sokal, 1973) is defined as a "...taxonomic character of two or more states, which within the study at hand can not be subdivided logically, except for subdivision brought about by the method of coding." This definition is particularly severe and not usually adhered to in practice. It is more often the custom to accumulate characters that are readily quantifiable, from different structures, with some care given to avoid the oversampling of individual structures to prevent the collection of highly correlated characters.

(2)The measures of similarity between two OTUs would be expected to stabilise about an equilibrium point as more characters are sampled. Although no specific work has been carried out through the random sampling of different character sets of varying sizes from a single large data set, this assumption is paramount, and is in fact implicit in most studies. The only feasible criticism of this is that when one has sampled compatible characters of one form, such as biochemical characters, it is conceivable that the inter-OTU relationships may be of a different form than if, say, morphological characters only were used. There are however strong empirical correlations between sampled characters of different forms. The pitfall at this stage is to sample only characters of one form and then, on that basis, to claim overall association. Until complete characterisation of the genome is possible through DNA sequencing, the ideal of a complete OTU as a stable numerical entity is unattainable.

(3)The lowest level of variability integrated into a study is inter-OTU not intra-OTU. To successfully identify groups, particularly OTU pairs, as distinct units, both the inter-group variability and the intra-group variability must be considered. Therefore, characters measured to typify an organism as an OTU must be stable with regard to the other OTUs selected. There are measures of association available that compensate for recorded intra-OTU variation (Sanghvi, 1953; and Crovello, 1968), but their integration with other characters that are stable, $S_i^2=0$, or non-quantitative, has not been recorded.

CHAPTER 2

RATIOS

"True love in this differs from gold and clay,
That to divide is not to take away."
PERCY BYSSHE SHELLEY

INTRODUCTION:

Given that a set of characters appropriate to the intended conclusions has been chosen, there frequently arises the contentious issue involving the construction of additional variables as a measure of shape, ratios, from the constituent set.

Atchley et al. (1976) highlighted the statistical inadequacies of using ratios as additional measures of plant form. They showed that ratios were generally right skewed and leptokurtic, increased the spurious correlations between variables, and did not, as traditionally believed, remove size effects and thus represent a shape variable. They argued from theoretical grounds, invoking the formulae of Chayes (1949) to show the randomly large correlations that occur between ratios and their constituent variables, and using simulations to show the effects of these correlations on Principal Components Analysis. Two years later, three responses to Atchley et al. (1976) were published in Systematic Zoology. Hills (1978) concurred with the original paper but showed how the use of log(x/y) lessened the undesirable properties of the untransformed ratio. Dodson (1978) argued from empirical grounds and while not disputing the value of the work of Atchley et al., he gave several examples where the use of ratios had helped to 'solve' a biological problem. Atchley replied to these responses in two further papers (Atchley, 1978; and Atchley & Anderson, 1978). Albrecht (1978) supported the points raised by Atchley et al. and by highlighting weaknesses in their methodology, further advanced their conclusions. Philips (1983) adopted a different approach to the 'ratio problem', in that while agreeing that ratios do not standardise for size he concentrated his attention on allometric growth patterns and showed that a non-linear change in the ratio with time adds significant 'error' to any so called definitive representation of shape. Clearly, in a numerical taxonomy context, where mature organs are being sampled, this criticism is not relevant.

Despite the general pertinence of the above criticisms, ratios have been and continue to be used to good effect in many 'real' numerical studies. Hill (1980), Estabrook & Gates (1984), West & Noble (1984) and Cantrill & Webb (1987) are but a few relevant examples.

To assess the actual impact of ratios in a numerical study, it is necessary to look initially at the definition of a 'useful' character and to examine how this relates to the ratio variable. It is by means of this orientation that characters are generally chosen to be measured and then potentially discarded if they add nothing to the final discrimination between groups. The description 'useful' incorporates two key points:

1)The character is adding new information to the data set. That is, it is not so highly correlated with other characters as to be merely adding spurious 'noise' to the study.

2)The character varies between taxa, so that a 'useful' character is "...any attribute of a member of a taxon by which it differs or may differ from a member of a different taxon" (Mayr, 1969).

Therefore, ratios may be calculated from continuous variables when appropriate, and then included in the study with no additional effort, until further manipulations elucidate their actual contribution.

METHODS:

Contained within the *Raoulia* data set are 21 ratio variables and their constituents. Of these, 17 represent some form of the shape of a given plant organ while the other 4 represent the ratio of the tubular and the filiform forms for particular floral structures. The form of the ratios, that is the choice of numerator and denominator (Appendix C), has been made by an experienced Numerical Taxonomist (Dr J.M. Ward) and thus unless specifically stated it is this form and not its reciprocal that is being studied. If any given ratio is inverted one is forming a different variable with different distributional characteristics. The normality of the ratio variables may be assessed by the univariate measures of skewness and kurtosis. Their impact on multivariate normality will be explored in the multivariate section of this thesis.

To assess the first premise of the 'useful' character statistically, it is necessary to measure the additional variability that a ratio adds to a data set. It is unlikely in any multivariate study of

adequate character and OTU size that the multiple coefficient of determination between any one variable and the remaining p-1 variables $R^2_{x.2\rightarrow p}$ will be less than about 0.80. Therefore, any tests associated with additional 'information' in regard to one character against all others will be unreasonably harsh. However if it can be shown that the 'residual' variability of a ratio, above and beyond that explained by its constituent variables, is significant, then one has a criterion by which to measure the additional usefulness of the ratio. Statistically this involves the calculating of the residual variability in the ratio, (z: z=x/y), after the effects of the constituents x and y have been removed, i.e., $1-R^2_{z.xy}$. Expressing the explained variation as the sum of the part correlation squared of z and y, ($R^2_{z(y.x)}$), and the coefficient of determination between z and x, ($R^2_{z.x}$), one has the 'explained' variation in z ($R^2_{z.yx} = R^2_{z(y.x)} + R^2_{z.x}$).

Given that

$$R^2_{z(y.x)} = \frac{(R_{z.y} - R_{z.x} * R_{y.x})^2}{1 - R^2_{y.x}}$$

the equation may be manipulated to the form

$$R^2_{z.yx} = \frac{R^2_{zy} + R^2_{zx} - 2 * R_{zy} * R_{zx} * R_{yx}}{1-R^2_{yx}}$$

Given this form we can look at the outcomes given biologically sensible correlation structures for R_{yx}, R_{zx} and R_{zy}. This technique will indicate what proportion of the variability in the ratio is redundant but will not in itself give any information on the usefulness of the additional variability. Invoking Mayr's (1969) description of a 'useful' character, and the inherent tendency among taxonomists to define useful characters as good diagnostic characters, one may quantitatively assess the 'usefulness' of a character. Given defined taxa, this may be done either by a one-way ANOVA, or allowing for the inter-character correlations by the standardised discriminant function coefficients. Campbell (1981) issues a note of caution to the effect that when there are significant correlations between the predictor variables (which is the case in this study) and this technique is being used to elucidate optimum grouping variables, the results will have to be interpreted cautiously. The standardised discriminant function coefficients may be calculated by using only the four variables discussed, that is, the two unit characters, their ratio and the ratio residuals. Any discriminating power contained in the residuals of the ratio will be contained in the

ratio, but by regressing out the effects of size we may obtain more precise discrimination, even if the residuals account for only a small proportion of the total variability in the ratio variable. The smaller this proportion, the less likely the ratio variable per se is to be useful.

From the variable list (Appendix C) three ratios (10, 79, 81) were chosen as exemplars of the above methodology; these represent three different but related structures: leaves, phyllaries, and capitula. The grouping variable to be used for testing the discriminating powers of the variables is based on the classification of Ward (1982) (Appendix A) with the exception of the single OTU species *R.cinerea*. The more general results on the distributions of the ratios will be extracted from the 17 'shape' ratio variables and the reciprocals of these ratios. The reciprocals can be used to test the hypothesis of Atchley et al. (1976) relating the degree of skewness and kurtosis in the ratio variable to the ratio of the coefficients of variation, $c.v._x/c.v._y$. Chayes' (1949) formulae for approximating the constituent-ratio correlations are calculated and compared to the actual values so the validity of using these simple calculations can be assessed.

RESULTS:

Table 1 gives the calculated proportion of variation in the ratio variable 'explained' by the constituents for different inter-correlation structures. The results for the different variables on the *a priori* groupings are given in Table 2, together with the inter-correlations, and the estimates of the inter-correlations as given by Chayes' (1949) formulae. Results from the correlations of all 17 shape ratio variables and their constituent variables show that the proportion of variability unique to the ratio combinations ranges from 0.01 to 0.47 with a mean of 0.20. The average inter-constituent correlation is 0.57, with a range of 0.10 to 0.96. The average correlation between the ratio and its numerator is 0.57 with a range of -0.06 to 0.95 and the average correlation between the denominator and its ratio is -0.22 with a range from -0.70 to 0.49. Table 3 gives the means, coefficients of variation, skewness and kurtosis measures and the ratio of the coefficients of variation for all 17 ratios and their reciprocals. Thirteen of these 17 ratios gave indication of a clear advantage to one form of the ratio over the other in terms of lower skewness and kurtosis measures. Within these 13 instances 10 of the favoured forms had ratios of the coefficients of variation greater than 1, and 8 had the ratio form with a mean less than or equal to 1. For the 34 measures of skewness and kurtosis for each ratio form, 14 were significant ($p<0.05$) when the c.v. ratio was greater than 1 and 25 when it was less than 1. A similar assessment when the ratio mean

TABLE 1: THE EXPLAINED VARIATION IN A RATIO (Z:X/Y) AS A RESULT OF X AND Y ($R^2_{z.xy}$) FOR DIFFERENT VARIABLE CORRELATION STRUCTURES

$R_{z.y}$ = 0.0

$R_{x.y}$ / $R_{z.x}$	0.0	0.4	0.8	0.9
0.0	0.0	0.0	0.0	0.0
0.2	.02	.05	.11	.21
0.4	.16	.19	.44	.84
0.6	.36	.43	1.0	
0.8	.64	.76		
0.9	.81	.96		

$R_{z.y}$ = −.4

$R_{x.y}$ / $R_{z.x}$	0.0	0.4	0.8	0.9
0.0	.16	.19	.44	.84
0.2	.20	.31	.91	
0.4	.32	.53		
0.6	.52	.85		
0.8	.80			
0.9	.97			

$R_{z.y}$ = −.6

$R_{x.y}$ / $R_{z.x}$	0.0	0.4	0.8	0.9
0.0	.36	.43	1.0	
0.2	.40	.59		
0.4	.52	.85		
0.6	.72			
0.8	1.0			
0.9				

$R_{z.y}$ = −.8

$R_{x.y}$ / $R_{z.x}$	0.0	0.4	0.8	0.9
0.0	.64	.76		
0.2	.68	.96		
0.4	.80			
0.6	1.0			
0.8				
0.9				

TABLE 2: PARAMETERS FOR THREE RATIOS AND THEIR CONSTITUENTS

		SDFC[1]	SDFC[2]	F-RATIO	CORREL.	CHAYES' APPROX.*
RATIO 1 (10)	Z	4.57	5.45	24.85	-.47[3]	-.49[3]
	Y	-	0.12	9.18	0.26[4]	-
	X	-.49	-.77	8.88	0.65[5]	0.68[5]
	RESID.	0.88	-	11.33		
		$R^2_{z.xy}$ =0.86				
RATIO 2 (79)	Z	-.46	-2.51	3.61	-.35[3]	-.32[3]
	Y	0.99	0.92	26.76	0.36[4]	-
	X	-	0.42	4.39	0.73[5]	0.74[5]
	RESID.	-2.1	-	5.23		
		$R^2_{z.xy}$ =0.96				
RATIO 3 (81)	Z	6.46	12.00	7.90	0.20[3]	0.18[3]
	Y	-	-3.45	19.85	0.77[4]	-
	X	-2.2	-3.16	27.35	0.77[5]	0.82[5]
	RESID.	5.53	-	4.34		
		$R^2_{z.xy}$ =0.97				

[1]: STANDARDISED DISCRIMINANT FUNCTION COEFFICIENTS

[2]: STANDARDISED DISCRIMINANT FUNCTION COEFFICIENTS WITHOUT THE RATIO RESIDUALS.

[3]: CORRELATION WITH Y.

[4]: CORRELATION WITH X.

[5]: CORRELATION WITH Z.

*:CHAYES' APPROXIMATION FORMULAE

$$r_{yz} = \frac{r_{xy}\,CV_x - CV_y}{(CV_x^2 + CV_y^2 - r_{xy}\,CV_x\,CV_y)^{1/2}}$$

$$r_{xz} = \frac{CV_x - r_{xy}\,CV_y}{(CV_x^2 + CV_y^2 - 2r_{xy}\,CV_x\,CV_y)^{1/2}}$$

TABLE 3: PARAMETERS FOR SHAPE RATIOS AND THEIR RECIPROCALS

	MEAN	C.V.	SKEW.	KURT.	C.V.x/C.V.y
1 Ratio (9)	0.61	.22	0.01	-.45	1.43
1/(9)	1.72	.25	1.20*+	1.47*	0.70
2 Ratio (10)	0.39	.37	0.37	-.85	1.18
1/(10)	2.57	.42	0.72*	-.46	0.84
3 Ratio (12)	0.81	.51	0.86*	0.90	0.83
1/(12)	1.63	.55	0.85*	-.30	1.21
4 Ratio (61)	0.07	.26	0.10+	-.90+	1.30
1/(61)	14.87	.28	0.76*+	-.29	0.77
5 Ratio (62)	0.06	.31	0.61*+	0.03	1.35
1/(62)	18.52	.33	0.83*+	0.23	0.74
6 Ratio (63)	0.11	.29	0.67*+	1.99*+	1.71
1/(63)	10.10	.33	1.39*+	2.31*+	0.58
7 Ratio (64)	0.07	.49	1.01*+	1.36*+	2.71
1/(64)	18.54	.52	1.20*+	1.10*+	0.37
8 Ratio (69)	0.39	.35	1.58*+	3.72*	0.81
1/(69)	2.79	.30	0.55*+	0.45	1.23
9 Ratio (70)	0.84	.48	1.55*+	2.05*+	0.70
1/(70)	1.42	.40	0.61*+	0.15	1.43
10 Ratio(73)	0.47	.33	1.16*+	1.97*+	0.92
1/(73)	2.34	.32	1.02*+	1.77*+	1.09
11 Ratio (74)	0.42	.25	0.87*+	1.13*+	0.96
1/(74)	2.50	.24	0.89*+	2.13*+	1.04
12 Ratio (76)	0.14	.24	0.40	-.21	1.74
1/(76)	7.40	.24	0.71*+	0.21	0.57
13 Ratio (77)	0.19	.17	0.49+	-.07	1.21
1/(77)	5.31	.17	0.28	-.39	0.83
14 Ratio (78)	0.20	.68	1.14*	0.53	3.70
1/(78)	7.51	.60	0.74*	-.22	0.27
15 Ratio (79)	0.19	.26	0.24	0.25	1.41
1/(79)	5.83	.31	1.48*+	2.57*+	0.71
16 Ratio (80)	1.00	.08	-.45+	0.05	0.97
1/(80)	1.01	.09	1.02*+	1.62*+	1.03
17 Ratio (81)	0.44	.25	-.29	-.29+	1.68
1/(81)	2.47	.33	1.72*+	2.73*+	0.60

* SIGNIFICANT P <0.05
\+ GREATER THAN RESPECTIVE VALUE FOR X & Y

was less than or equal to 1 gave 17 significant results as against 22 when the mean was greater than 1. No analysis has been carried out on the grouping ability of the reciprocals of the ratios, and it is almost certain that these differ from those of the ratios used here. Further work is clearly required here to see if any association exists between the discriminating ability of the ratio and either the mean of the ratio (greater than 1 or less than 1) or the univariate skewness and kurtosis of the ratio.

DISCUSSION:

It is evident from Table 2, and from the calculations on the entire ratio set, that significant correlations occur within the general ratio structure when genuine biological data is being used. Tables 1 and 2 give some indication of the effects of these high correlations on the proportion of variability unique to the ratio variable. The lower magnitude of the denominator to ratio correlation over the numerator to ratio correlation is likely to be the product of the generally smaller values of the denominator coefficient of variation. The average inter-constituent correlation of 0.57 is possibly less than might have been expected, but the range for this indicates that it can become very large. The overall effect of the three average correlations is to make the proportion of 'non-size' dependent variability in the ratio variable approximately 0.20. Clearly by including all three variables numerator, denominator and the ratio in any analysis, at least approximately 1/4 of the combined 'information' is redundant. This is likely to be considered unacceptably high. Working from the assumption that a maximum amount of retained variability is an optimum solution, regardless of the 'quality' of this variability, one clearly has three options in terms of removing a single variable. Using the three averages for R_{zx}, R_{zy} and R_{xy} one can easily compute the amount of variability redundant to each of the three variables. These are 0.76 for both the ratio and the denominator and 0.83 for the numerator. The difference is clearly not great but in this instance it does indicate that on average a better strategy than not using a ratio is to create it and then remove the numerator from the data set. This result need not be taken as an empirical generality but can be easily assessed in any particular instance. For the 17 ratio variables in this study this strategy would be appropriate in 11 cases.

The results from Table 2 indicate that Chayes' (1949) approximations for R_{zy} and R_{zx} are reliable enough and therefore their use to approximate correlations, in situations such as that described where particular strategies are being compared, is vindicated. Atchley et al. (loc. cit.)

claimed that the approximation gave a consistent overestimate for R_{zx}, and this appears to be verified in these results. The absolute error in the approximation may well increase with the size of the estimate but the percentage error looks likely to maximise at low magnitudes. Based on just the three examples given here, this trend can be stated only as a likelihood. The use of 'real' data in these examples meant the ratio of the coefficients of variation varied from 1.2-1.7, so generalisations beyond this can not be made.

The usefulness of the variability in each variable was considered in two ways (Table 2): one assuming variable independence- one way-anova, and the other compensating for variable colinearity, the relative magnitudes of the standardised discriminant function coefficients. The first measure (F-ratio) indicates that any of the four variables involved in a ratio, would provide significant independent discrimination for any of the three ratios ($p<.001$). The rankings within each combination indicate no clear advantage in using any one variable. Ignoring the residual variable this result is to be expected, but the 'success' of the residual variable in terms of its discriminating power is a surprise. The F-ratios for the ratios give no indication that they have consistently less discriminating power than either of their constituents. In fact when the 'residual' variable is excluded and thus its discriminating power within the discriminant function is amalgamated with the ratio variable, then the ratio emerges as the most highly correlated variable with the grouping variable in all three examples. On the basis of this result one can deduce that of the three variables describing the three largely independent structures, the ratio is the most useful in each instance.

The results from Table 3 indicate that there is a strong relationship (11/17) between forming a ratio so that the ratio of the coefficients of variation is greater than 1 and forming one that has a mean less than 1. This implies that differences between the means of two size variables on a given structure are not necessarily accompanied by a proportionate change in standard deviations. The skewness and kurtosis measures from Table 3 indicate that a clear advantage can be achieved by creating the ratio so that the ratio of the c.v.s is greater than 1. A correlated but less pronounced advantage can be achieved if the ratio is formed so that its mean is less than 1. Although an advantage is seen in either of these policies, it is apparent that considerable non-normality still occurs regardless of the form of the ratio. In general the ratios are right-skewed and leptokurtic when they are not normally distributed. This appears to conflict to some degree with

the general findings of Atchley et. al. (loc. cit.). The results from which they generated their conclusions were based on computer generated data, where the parametric means were all set to a single constant, and samples were generated from this for predefined coefficients of variation etc. They showed that when the ratio of the coefficients of variation was approximately equal to 1 and increasing, both the skewness and the kurtosis of the ratio were stabilising about a minimum value. Empirically it would seem that a ratio of the coefficients of variation in a real example is most likely to be approximately 1, since we have no reason to believe that the two constituents measured on the same scale should have significantly different mean-standard deviation ratios. Furthermore, in the examples of Atchley et al. (loc. cit.) all the manipulations were based on ratios formed from two constituents with identical parametric means, which is biologically implausible. The use of two variables with identical population means to create a 'shape' variable makes little biological sense. It appears likely that the creation of ratios as depicted by Atchley et. al. (loc. lit.) does lead to ratios that have non-normal distributions. When, however, the ratios of the coefficients of variation are approximately equal and the ratio is created so that this ratio is greater than 1, then it would seem likely that the resultant ratio variable is less likely to show any of the extreme non-normality shown by Atchley et al. (loc. cit.). The degree to which the existing non-normality is of concern depends primarily on the distributions of the other variables within the data set. It may be that particular outliers within the set are tending to make a great number of the variables non-normally distributed. The 17 ratios in the study were all formed to have a mean less than 1. Of these, 6 show skewness and kurtosis greater than both constituents. If 6 of these ratios were reformed so that all ratios had c.v. ratios greater than 1, then again 6 show non-normality greater than both constituents. If any three variables were randomly selected from a data set then one might expect, with a probability of .11, that any one of them would have consistently greater non-normality in two measures. The results here suggest about a three-fold bias towards the ratio variables, created by either criterion, being the most non-normal in both measures. Indications from Table 3 are that this bias would be slightly higher if one was to consider only significant skewness and kurtosis. Whether this increase in the number of non-normal univariate distributions has significant repercussions is debatable. In most studies the distributions of variables are analysed before any form of grouping has been ascertained. The reason they are analysed is to see whether there are any violations of assumptions for parametric statistics. However the major assumptions for the multivariate studies of numerical taxonomy involve multivariate normality and homogeneity

of the variance-covariance matrices. The influence of individual variable distributions on this will be explored in a subsequent chapter; it is sufficient to say here that it is frequently the skewed variables which provide the best non-parametric discrimination of a single sub-group from the remaining cases.

The coefficients of variation for the ratios themselves, while not within the range said to be acceptable for biological data, are within the ranges determined by their constituents. The choice of numerator and denominator appears to influence the size of the coefficient of variation. By forming all ratios so that the means are less than or equal to 1 then 12 of the 17 c.v.s are optimised whereas creating the ratios so that the ratio of the c.v.s is greater than 1 optimises 14 of the c.v.s. The ranges of the c.v.s for the variables within the sub-genus being studied show that there is considerable inter-case variability, indicating the likely presence of smaller sub-groups within the subgenus.

CONCLUSIONS:

It has been shown that significant correlations occur both between components chosen to represent the 'shape' of given plant structures and between these two constituents and the 'shape' variable of their combination. These correlations lead to a significant amount of variability being included more than once into a data set if all three variables are entered. A simple formula has been constructed that enables the loss of information to be minimised when one of the three variables is omitted. A general recommendation for the omission of the numerator rather than the ratio is made, given that the ratio is constructed on the basis of either of two criteria. The 'usefulness' of the ratio variable has been assessed by splitting off the 'non-size' dependent component (residual), and then in so doing showing that this component offers significant discriminating ability between the 13 *a priori* groups. In the standardised discriminant function, the ratio itself provides the best discrimination in each example. The skewness, kurtosis and the c.v.s for the ratios and their reciprocals were analysed. This shows that by creating the ratio variable so that the numerator corresponded to the variable with the higher c.v., in general one would be minimising the non-normality of the ratio *per se*. A slightly inferior policy was to create the ratio so that the numerator had the smaller mean. Significant non-normality was still shown to be present among the ratios studied but the implications of this are not considered serious as they stand. Further work on the multivariate distribution will elucidate their full impact. There is

obviously further work still to be done on the influence of numerator-denominator choice on both the discriminating ability of the ratio, and the proportion of redundant information in each of the three variables.

CHAPTER 3

MULTIVARIATE NORMALITY

"For there is a music wherever there is a harmony, order or proportion; and thus far we may maintain the music of the [hyper-] spheres; for those well ordered motions, and regular paces, though they give no sound unto the ear, yet to the understanding they strike a note most full of harmony." THOMAS BROWNE

INTRODUCTION:

Multivariate normality is an explicit assumption for the probabilistic interpretation of results for a number of multivariate statistical techniques such as, for example, discriminant analysis. Despite this, none of the major statistical packages and very few exponents of the techniques validate their results by checking this assumption. Three notable exceptions to this are Campbell and Mahon (1974), Reyment (1971), and Birks and Peglar (1980).

There are three likely reasons for this neglect:

(i)The lack of practical application by the theoretical statisticians who developed the methodology, particularly R.A. Fisher and K.V. Mardia, even in collaboration with practising researchers, which makes the techniques for testing this assumption beyond the scope of the non-statistical practitioners who should be employing them.

(ii)The blithe dismissal of any concern by some scientists, e.g., Green (1971) by stating that the data is 'likely' to be multivariate normal.

(iii)The application of the techniques by authors who correctly note that the method, specifically discriminant analysis, does not require the assumption *per se*, except when probabilities are to be accurately quoted and interpreted (Pimental, 1981; Campbell & Mahon, 1974).

The first two of these points necessitate that the practical implications of non-normality be further explored, particularly in situations in which the number of variables is greater than 4. There are often sound empirical reasons for expecting a sample from a population to be

multivariate normal, and any departures from this are likely to be of practical interest in terms of both the variables measured and the OTUs sampled.

The theoretical work of Mardia(1970, 1974, 1975) provides the appropriate framework from which to assess the normality, via the multivariate measures of skewness and kurtosis, of any multivariate data set. Mardia(1970) has derived the following as sample measures of multivariate skewness and kurtosis respectively:

$$b_{1,P} = 1/n^2 \ \Sigma \ \Sigma \{(X_i-X)'S^{-1}(X_j-X)\}^3$$

$$b_{2,P} = 1/n \ \ \Sigma \{(X_i-X)'S^{-1}(X_i-X)\}^2$$

for a set of n OTUs measured on p variables $X_1,....X_n$, and where X denotes the sample mean vector and S^{-1} the inverse of the sample variance-covariance matrix.

For the general case Mardia(1974) has shown that:

$$E(b_{1,P}) = p(p+2)\{(n+1)(p+1)-6\} / \{(n+1)(n+3)\}$$

and from this he generates three methods for assessing the degree of skewness via the Chi-square and Normal distributions:

(i)A is distributed as χ^2 with $p(p+1)(p+2)/6$ d.f where $A = nb_{1,P}/6$

(ii)A' is distributed as χ^2 with $p(p+1)(p+2)/6$ d.f where $A' = knb_{1,P}/6$ and

$k = (p+1)(n+1)(n+3)/[n\{(n+1)(p+1)-6\}]$ for all n.

iii)For $p>7$ using the normal approximation to the Chi-square distribution for (i): $(2A)^{1/2}$ is distributed as $N[\{(P(P+1)(P+2)-3)/3\}^{1/2},1]$.

He further shows that:

$$E(b_{2,P}) = p(p+2)(n-1)/(n+1) \text{ with}$$

$$Var(b_{2,P}) = C\{8p(p+2)(n-3)\}/\{(n+1)^2(n+3)(n+5)\}$$

$$\text{where } C = (n-p-1)(n-p+1)$$

From this he generates two normal approximations for testing departures from normality in terms of multivariate kurtosis:

(iv)B' is asymptotically distributed as N(0,1) where

$$B' = \{(n+1)b_{2,P} - p(p+2)(n-1)\}\{(n+3)(n+5)\}^{1/2}$$
$$\{8p(p+2)(n-3)(n-p-1)(n-p+1)\}^{1/2}$$

(v)For large n, B is asymptotically distributed as N(0,1) where

$$B = \{b_{2,P} - p(p+2)\}/\{8p(p+2)/n\}^{1/2}$$

A corollary to the calculation and testing of $b_{2,P}$, is the fact that the kurtosis measure is the average Mahalanobis D^2 squared of each OTU to the centroid ie $b_{2,P} = 1/n \Sigma (D_i^2)^2$, which then enables one to test any individual D_i^2 as an approximation to the χ^2 distribution with p d.f.

Reyment (1971) used formulae (i) and (iv) to test for significant skewness and kurtosis respectively in a number of large data sets all with $p <= 4$. From these results he showed the lack of correlation between univariate measures of skewness and kurtosis and their multivariate equivalents. He further showed, by taking subsets from these data sets, that any sample measure of multivariate skewness tends to stabilise with a sample size of approximately 100, and that this stability may occur earlier when the number of variables is increased. While applying no comparable kurtosis tests for differing variable numbers, he did show that kurtosis values do not stabilise as quickly as skewness values for $p=4$. These results invite more questions than they answer, particularly in relation to the influence of individual variables and variable number on overall measures of multivariate normality. Mardia (1975) briefly summarised the theoretical interpretation of skewness and kurtosis anomalies in terms of data points (OTUs). He indicated that significant skewness implies an imbalance or excessive clustering of points around the multivariate centroid, and significant kurtosis indicates an imbalance in the radial distribution of the points. Generally then, skewness distortions can be said to correspond to shape distortions, whereas significant kurtosis measures are associated with size distortions.

METHODS:

To assess the sample measures of multivariate skewness and kurtosis for the *Raoulia* data set, a FORTRAN program (Appendix D) was written which used the algorithms of Healy (1968a, 1968b) and Mardia & Zemroch (1975). In addition to this, the program was written to give the Normal and Chi-square test statistics for the estimates and also the univariate measures of skewness and kurtosis for each variable. The individual Mahalanobis D_i^2s from each OTU to the group

centroid are also calculated; these may be plotted in the manner of a normal probability plot (Healy, 1968c) in order to assess the multivariate normality and to detect outliers. With the above results it is possible to ascertain the individual variable/OTU effects on the overall distribution, by selecting particular OTU/variable subsets for separate analysis. Only the continuous variables in the data set were used ($p=63$) in analysis 1. Given the asymptotic nature of some of the test statistics (Mardia, 1970; Hawkins, 1974), and the fact that no multivariate assessment of normality with $p>10$ appears to have been attempted, the values of the test statistics will tend to be interpreted in a relative rather than an absolute manner.

Cluster analysis was employed to relate the findings of the above methodology (analysis 1) specifically in relation to OTU outliers, with those of an accepted technique. In this manner the effect of outliers on the general associations portrayed by cluster analysis may be seen, and the 'arbitrary' affinity ascribed to them if they are not shown as outliers revealed. The cluster analysis used here employed the 'range coefficient' integral to Gower's coefficient (Gower, 1971) as the measure of association and the UWPGMAA (Sneath & Sokal, 1973) as a clustering technique. These specific options were chosen to represent commonly accepted and recommended techniques (Sneath & Sokal, 1973).

The impact of the ratios in the data set on the overall multivariate normality will be gauged by removing all the ratios and recalculating the multivariate measures of skewness and kurtosis, with $p=37$. To test specifically if the 'shape' ratios themselves were an integral component of the multivariate normal distribution of the sub-genus, only these ratios were included, and again the multivariate measures of skewness and kurtosis were calculated ($p=54$). This latter step involved the exclusion of three shape characters whose components were not represented in the data set.

To explore the impact of normality/non-normality on a two-group discriminant analysis the OTU set will be bisected using three independent criteria (analysis 2). For a complete description of the methodology of 2-group discriminant analysis see Lachenbruch (1975) and the subsequent chapter of this thesis. The criteria for bisection are *Raoulia hookeri*/not *Raoulia hookeri* (Allan, 1961), diploid/polyploid and riverbed inhabitants/non-riverbed inhabitants. Using these dichotomies, the individual multivariate normalities of the six groups may be assessed and then related in a practical manner to the findings of the three pair-wise discriminant analyses. A

significant disadvantage to numerical taxonomists in the calculation of the multivariate measures of skewness and kurtosis is the fact that the number of variables used cannot exceed the number of OTUs. In subdividing the data set into two approximately equal subsets, this problem automatically occurs. To overcome this, drastic pruning of the variables, down to p=29, was necessary. At this stage in the analysis no criteria were being placed on the variables in terms of discriminating ability or distributions. Therefore, the best stipulate by which to reduce the variable number was considered to be that of bivariate collinearity, i.e., to remove those variables most highly correlated, bivariately, with the other variables. For this purpose a method was employed whereby variables were removed in a stepwise manner as they maximally reduced the number of 'significant' bivariate correlations. A case could be made for selecting variables in a manner completely opposite to this method, i.e., to leave an agglomeration of highly correlated characters as a core for the subsequent analysis. However, if both multivariate normal and non-normal distributions can be generated then the selection technique will be of little consequence. The major disadvantage in having to create such a subset is that it impairs the comparative assessment between results from the single distribution and those in which the OTUs are subdivided into two groups.

RESULTS:

Analysis 1: The multivariate measures of skewness and kurtosis for the full data set (n=86, p=63) were 3107.7 and 4073.11 respectively. The z-scores associated with these estimates were 2.91 and 3.70, both significant at $p<.01$. Although these measures do not indicate extreme non-normality, they do indicate the existence of an irregular multivariate distribution. To see if this distribution was 'improved' if univariate normality existed for all variables independently, the 24 variables that showed significant univariate skewness or kurtosis ($p<.01$) were removed and the analysis re-run. Of the 24 variables removed, 23 displayed significant ($p<.05$) non-normality in terms of both skewness and kurtosis, giving an indication of the correlation that exists between these two measures in genuine biological samples of this type. The 24 variables removed included:

(i) All leaf measurements except length of leaf base, breadth of leaf base, shoot diameter, and lamina length/leaf length ratio.

(ii) All floral counts except phyllary number and percentage of female florets.

(iii) Length/maximum breadth of phyllary ratio.

(iv) eight floral ratios involving achene and corolla measurements.

The resultant measures of multivariate skewness and kurtosis for this reduced data set were 979.1 and 1711.6 with z-scores of 21.6 and 12.2 respectively. Clearly the normality of the multivariate distribution was further distorted rather than improved by the deletion of the non-normal variables. Consequently it was decided that the influence of OTU outliers on the multivariate distribution should be explored. This was achieved by removing the 6 OTU outliers that showed significant ($p<0.05$, $\chi^2=82.2$) Mahalanobis D^2s from the multivariate centroid of the subgenus, (Table 4). It should be noted that given $n=86$ one would expect about 4 outliers as a result of random sampling using $\alpha=0.05$, so in using this α level a rigid criterion for non-outliers has been set. These six 'outlying' OTUs included two OTUs from *R.parkii* (1,6), one from *R.hookeri h.* (30), one from *R.hookeri as.* (34), one from *R.hookeri an.* (37) and the single species *R.cinerea* (77). The measures of skewness and kurtosis for this reduced data set ($n=80$, $p=63$) were 2950.2 and 3784.4 with z-scores of -1.33 and 1.85 respectively. Thus the deletion of the OTU outliers, as defined by an excessive Mahalanobis D^2 from the OTU to the multivariate centroid, leads in this instance to a multivariate normal distribution as defined by the measures of skewness and kurtosis. The effect of the removal of the six OTUs on the univariate distributions was minimal. 21 rather than 24 variables now had significant ($p<0.01$), measures for skewness and kurtosis, with the major impact being on the leaf measurements where the skewness and kurtosis were lowered dramatically. This would seem to reflect the larger vegetative size of some of these outliers. The technique of removing outliers and then reassessing the univariate distributions provides an effective tool for identifying those characters that best differentiate the outliers.

The cluster analysis based on the 63 continuous characters represents the assumed classification (Ward 1982) quite well (Fig.1) and thus little information in the form of diagnostic discrete characters is missing from this reduced data set. At first glance three OTUs stand out as being potential outliers. Of these, two (*R. cinerea* (77) & *R. hookeri as.* (34)), were revealed in the earlier analysis, the one that did not appear earlier being *R.'sp.K'* (74). Proceeding to extract the six (7) most weakly aligned individual OTUs reveals that three of these were highlighted in the previous analysis, the additional one being *R.hookeri h.* (30). The remaining four; (i)*R.'sp.K'* (74) (ii)*R.'sp.K'×tenuic.*(76) (iii)*R.'sp M'* (57) and (iv)*R.hookeri h.* (32), can be explained in terms of:

(a)Their 'random' associations with verified outliers *R.hookeri h.* (32),

TABLE 4: MAHALANOBIS D^2 S OF ALL OTUS TO THE CENTROID OF THE SUBGENUS

		D^2			D^2
OTU	1	84.5*	OTU	44	43.8
OTU	2	63.0	OTU	45	56.2
OTU	3	64.4	OTU	46	55.9
OTU	4	55.7	OTU	47	69.0
OTU	5	65.2	OTU	48	57.6
OTU	6	82.4*	OTU	49	72.5
OTU	7	58.5	OTU	50	74.6
OTU	8	75.6	OTU	51	71.8
OTU	9	61.7	OTU	52	67.3
OTU	10	65.8	OTU	53	61.2
OTU	11	67.6	OTU	54	56.1
OTU	12	63.7	OTU	55	55.0
OTU	13	63.9	OTU	56	51.3
OTU	14	57.5	OTU	57	65.7
OTU	15	69.4	OTU	58	72.8
OTU	16	48.2	OTU	59	73.6
OTU	17	51.0	OTU	60	71.4
OTU	18	44.3	OTU	61	60.3
OTU	19	62.4	OTU	62	72.8
OTU	20	65.6	OTU	63	64.7
OTU	21	56.0	OTU	64	56.2
OTU	22	58.2	OTU	65	75.9
OTU	23	57.4	OTU	66	57.3
OTU	24	62.8	OTU	67	53.3
OTU	25	63.9	OTU	68	57.2
OTU	26	46.2	OTU	69	69.1
OTU	27	68.3	OTU	70	51.4
OTU	28	69.3	OTU	71	69.3
OTU	29	69.4	OTU	72	75.3
OTU	30	83.7*	OTU	73	76.9
OTU	31	51.9	OTU	74	59.6
OTU	32	71.3	OTU	75	78.3
OTU	33	50.8	OTU	76	71.8
OTU	34	82.9*	OTU	77	82.4*
OTU	35	56.7	OTU	78	67.8
OTU	36	44.8	OTU	79	68.3
OTU	37	85.0*	OTU	80	57.9
OTU	38	55.4	OTU	81	54.4
OTU	39	47.8	OTU	82	67.1
OTU	40	44.9	OTU	83	63.6
OTU	41	60.3	OTU	84	68.8
OTU	42	51.3	OTU	85	64.1
OTU	43	40.7	OTU	86	53.5

*OTU BEYOND 5% LIMIT FOR NORMAL DISTRIBUTION

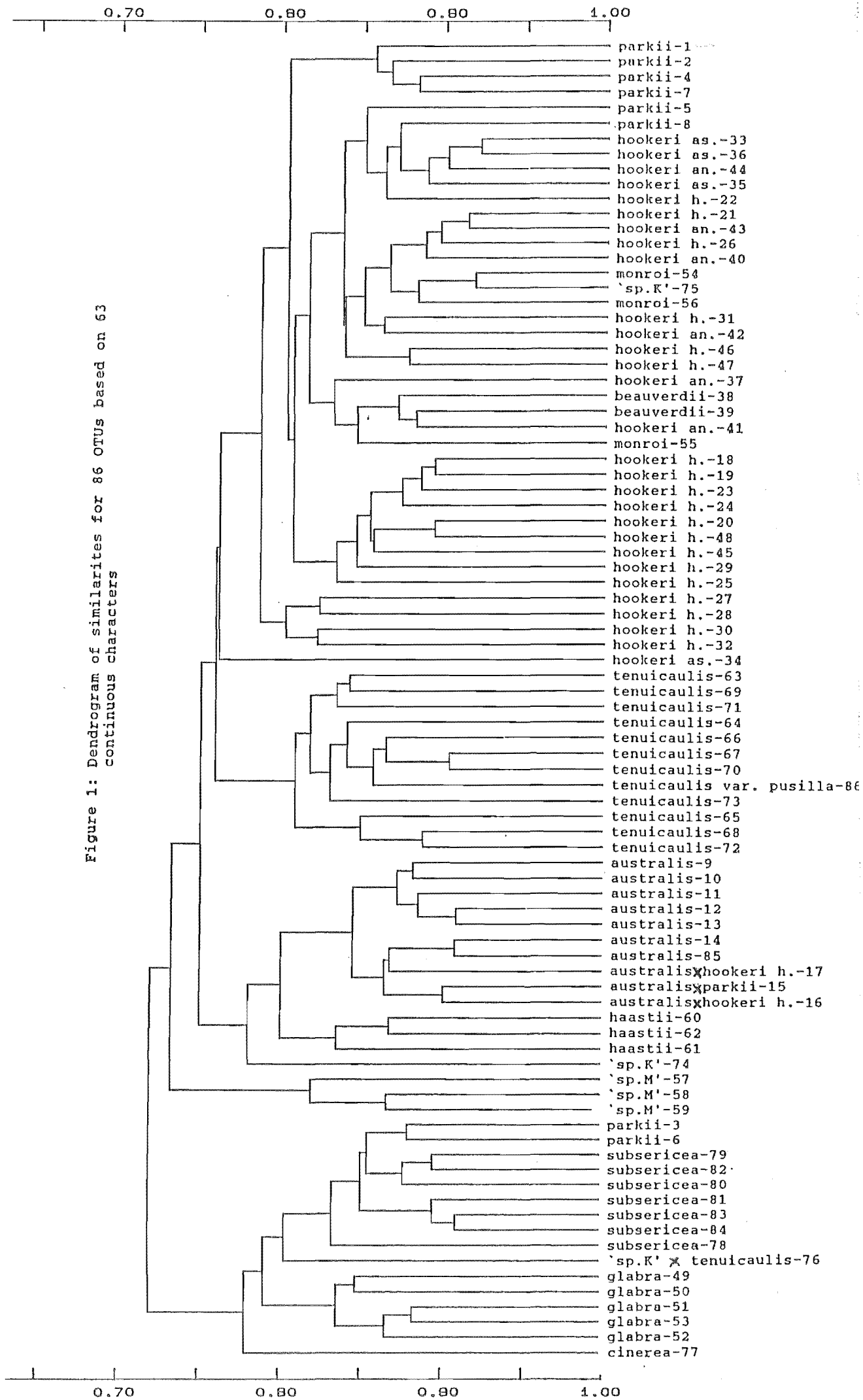

Figure 1: Dendrogram of similarites for 86 OTUs based on 63 continuous characters

(b)The loose nature of their postulated species *R.'sp M'* (57), and

(c)The indefinite relationship between their proposed species *R.'sp.K'* (74) and *R.'sp.K'* & *tenuic.*(76), (the remaining member of *R.'sp.K'* (75) was only marginally below the critical χ^2 value for the D^2 assessment of outliers, (Table 4), and it showed a 'random' but strong affiliation for an OTU of a different species here, thus potentially 'spoiling' the clustering of that species as well).

The three OTUs revealed as outliers in the previous analysis but not shown up in the cluster analysis all reveal slightly unusual associations, with *R.parkii* (1) and *R.hookeri an.* (37), being on the extremity of their clusters, and *R.parkii* (6) associating with *R.parkii* (3) a considerable distance from the *'parkii'* cluster. The removal of the six outliers had varying repercussions within the phenogram (Fig.2), which depended largely on the degree of abnormality they portrayed in the previous cluster analysis. The removal of the two extreme OTUs *R.cinerea* and *R.hookeri as.* (34) had no effect on the clusterings of the other OTUs. The extraction of the two *R.parkii* OTUs (1,6) however, confirmed the apparent split of the *'parkii'* group into a distinct cluster and one that was inextricably combined with the *'hookeri'* complex. Obviously *R.parkii* (1) was largely responsible for giving the one cluster a distinct appearance and *R.parkii* (6) was giving the impression that *R.parkii* (3) was also a very divergent *'parkii'*. The removal of *R.hookeri h.* (30) and *R.hookeri h.* (37) caused the repositioning of their individually most closely associated OTUs. The absence of *R.hookeri h.* (30) caused *R.hookeri h.* (32) to be moved from the tight *R.hookeri h.* cluster and the removal of *R.hookeri an.* (37) exacerbated the removal of *R.parkii* (1) by causing *R.beauverdii* (38), *R.beauverdii* (39), *R.hookeri an.* (41), and *R.monroi* (55) to be positioned between the distinct *'parkii'* group and the apparent *'hookeri-parkii'* complex.

The exclusion of all the ratios yielded z-scores for the measures of multivariate skewness and kurtosis of 18.4 and 10.7 respectively. When the 'shape' ratios were added to the set these values fell to 7.4 and 5.4.

Analysis 2: i) The subdivision of the data set into the *R.hookeri* OTUs (n=33) and the *'non-hookeri'* OTUs (n=53) yielded z-scores for the multivariate measures of skewness and kurtosis of 4.07, 0.10 and 0.39, 1.11 respectively. The large skewness measure for the *R.hookeri* subset possibly implies a shape distortion, multi-modes, rather than a few outlying OTUs. A detailed scan

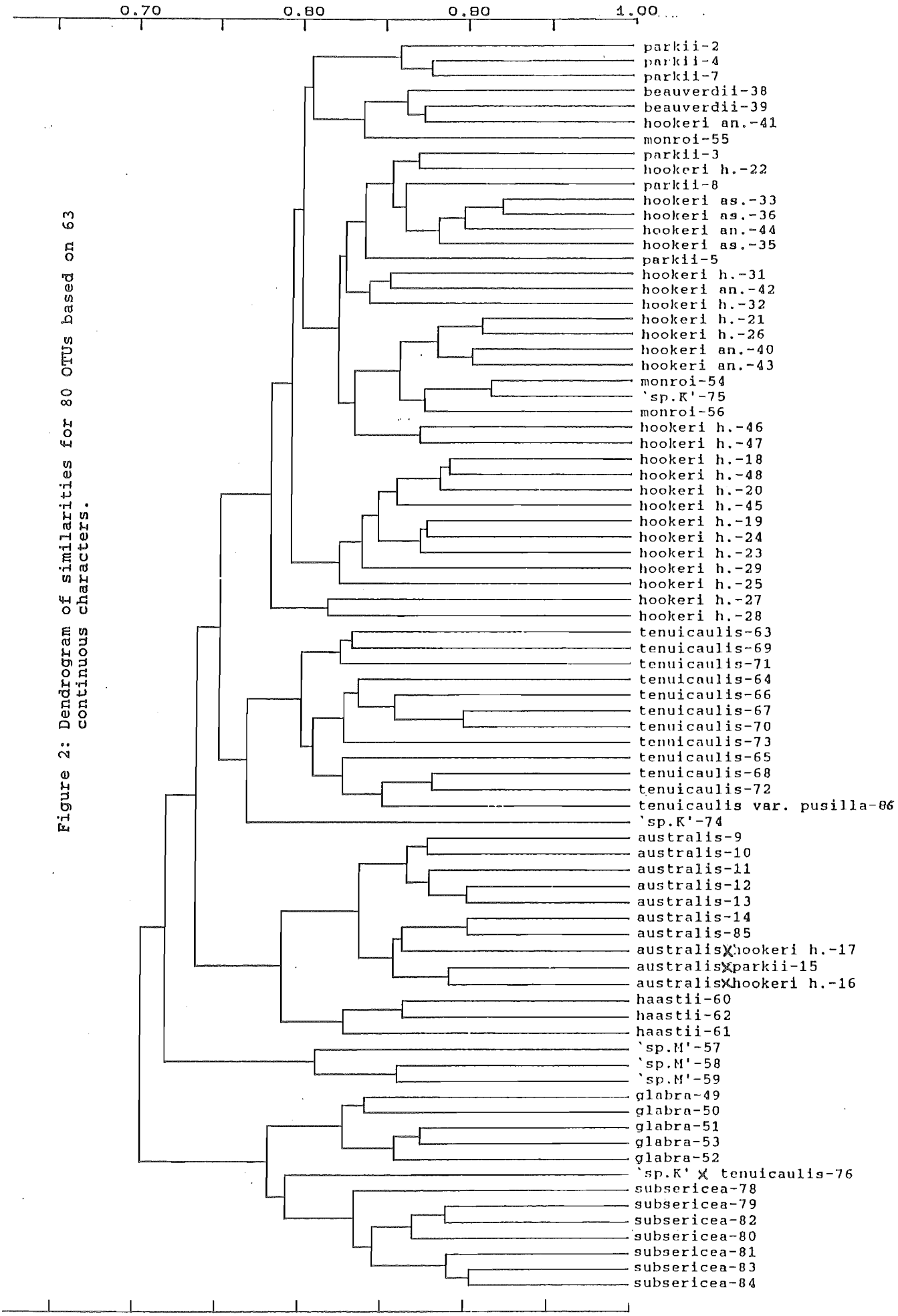

Figure 2: Dendrogram of similarities for 80 OTUs based on 63 continuous characters.

of the D^2 values for the 33 OTUs confirms this, with none outside the χ^2 ($p<0.05$) limit of 40.0 (Table 5). Four of the variables exhibited grossly non-normal patterns, and, as might be expected, the individual OTUs responsible for this were not consistent through the four variables. The two hybrids (OTUs 16 & 17) included in this group showed no strong tendency to be aligned outside the group, despite the fact that OTU 16 did have the largest mahalanobis D^2 of 31.6 (Table 5), a value that falls well short of the critical value of 40.0 at $p < 0.05$. The *'non-hookeri'* OTUs are shown by the multivariate measures of skewness and kurtosis to be a more 'normal' group than the *R.hookeri* species itself. This fact is reinforced by the existence of only one variable showing a non-normal univariate distribution. The Mahalanobis D^2s, however, show considerably more variation, 17.1-46.1 as against 19.7-31.6, (Table 5), when compared with the *R.hookeri* group. The value of 46.1 for OTU 77 easily exceeds the critical value at $p<0.05$. This is not surprising, as this OTU (*R.cinerea*) is considered the least *Raoulia*-like species in the entire subgenus (Ward, 1982). The canonical variate analysis using all 29 variables gives significant separation between the two centroids ($p<0.01$) but considerable overlap was still present (Fig.3), with 12 misclassifications on the basis of a minimum D^2 from each OTU to either of the centroids. None of the 12 misclassifications correlated with high D^2s (centroid to OTU) from the earlier multivariate normal assessment. None of the five variables that showed non-normal univariate distributions in the previous study featured prominently in the list of standardised canonical function coefficients. OTU 77, considered an outlier from the *'non-hookeri'* group, had the largest mahalanobis D^2s from the two centroids of 60.6 and 67.9 but was not misclassified.

(ii) The second subdivision of the data set into diploid OTUs ($n=31$) and polyploid OTUs ($n=54$) gave z-scores for the multivariate measures of skewness and kurtosis of 4.64, 0.04 and 0.91, 1.51 respectively, (*R.cinerea* was not included in this analysis as no chromosone count was available). The aberrant skewness measure for the diploid group may be a product of the lack of univariate normality with four variables (all floral counts) which showed extreme non-normality. The few OTUs causing these unusual distributions are not consistent through the four variables, and, as a consequence, the D^2s form a homogenous subset ranging from 25.5-30.0 (Table 6), with none beyond the critical χ^2 value of 40.0 at $p<0.05$. The polyploid subset appears to form a more homogenous group in terms of the multivariate measures of skewness and kurtosis. However, four different variables show significantly non-normal distributions while the range of the D^2s is larger,

TABLE 5: MAHALANOBIS D^2S FOR *HOOKERI* AND NON-*HOOKERI* GROUPS

HOOKERI			NON-*HOOKERI*					
OTU	16	31.6	OTU	1	31.6	OTU	61	32.2
OTU	17	27.5	OTU	2	28.1	OTU	62	35.6
OTU	18	29.9	OTU	3	25.1	OTU	63	29.5
OTU	19	31.5	OTU	4	22.9	OTU	64	27.4
OTU	20	27.6	OTU	5	32.5	OTU	65	35.0
OTU	21	25.5	OTU	6	22.7	OTU	66	25.6
OTU	22	27.5	OTU	7	28.9	OTU	67	23.8
OTU	23	31.6	OTU	8	20.3	OTU	68	17.8
OTU	24	31.0	OTU	9	30.7	OTU	69	31.2
OTU	25	31.1	OTU	10	33.6	OTU	70	21.4
OTU	26	25.3	OTU	11	25.1	OTU	71	37.7
OTU	27	31.1	OTU	12	35.4	OTU	72	25.7
OTU	28	30.1	OTU	13	31.8	OTU	73	28.0
OTU	29	30.5	OTU	14	18.0	OTU	74	34.7
OTU	30	31.4	OTU	15	33.3	OTU	75	31.6
OTU	31	25.0	OTU	49	36.9	OTU	76	28.3
OTU	32	30.2	OTU	50	29.3	OTU	77	46.1*
OTU	33	29.7	OTU	51	30.8	OTU	78	24.5
OTU	34	26.4	OTU	52	40.4*	OTU	79	31.5
OTU	35	28.4	OTU	53	24.0	OTU	80	22.0
OTU	36	19.7	OTU	54	17.1	OTU	81	25.6
OTU	37	30.7	OTU	55	29.0	OTU	82	26.9
OTU	38	26.7	OTU	56	28.7	OTU	83	17.8
OTU	39	30.9	OTU	57	37.4	OTU	84	27.8
OTU	40	29.6	OTU	58	35.0	OTU	85	22.8
OTU	41	29.8	OTU	59	34.7	OTU	86	26.0
OTU	42	31.1	OTU	60	37.4			
OTU	43	30.9						
OTU	44	29.3						
OTU	45	28.7						
OTU	46	28.4						
OTU	47	31.1						
OTU	48	27.2						

*OTU BEYOND 5% LIMIT FOR NORMAL DISTRIBUTION

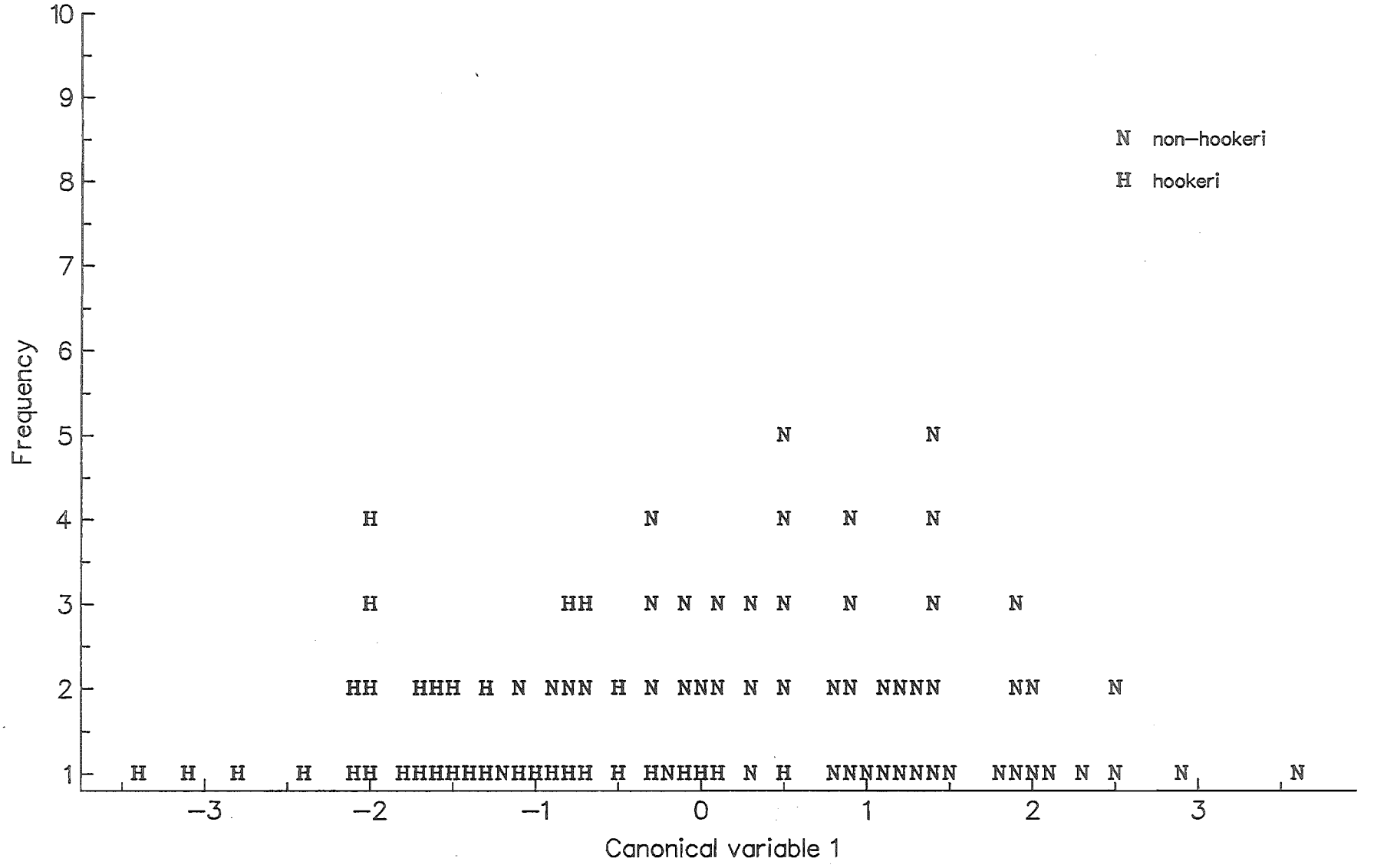

Figure 3: Plot of canonical variable 1 for hookeri and non–hookeri groups.

TABLE 6: MAHALANOBIS D^2S FOR DIPLOID AND NON-DIPLOID GROUPS

DIPLOID			NON-DIPLOID					
OTU	9	28.4	OTU	1	37.1	OTU	33	23.5
OTU	10	29.5	OTU	2	28.0	OTU	34	31.4
OTU	11	29.5	OTU	3	29.9	OTU	35	29.4
OTU	12	29.9	OTU	4	29.2	OTU	36	23.1
OTU	13	29.8	OTU	5	35.7	OTU	37	43.5*
OTU	49	30.0	OTU	6	26.0	OTU	38	27.8
OTU	50	28.7	OTU	7	25.1	OTU	39	27.9
OTU	51	27.9	OTU	8	27.5	OTU	40	19.8
OTU	52	27.8	OTU	14	23.4	OTU	41	26.5
OTU	53	27.1	OTU	15	37.2	OTU	42	21.1
OTU	54	29.3	OTU	16	31.9	OTU	43	22.6
OTU	55	30.0	OTU	17	21.1	OTU	44	21.3
OTU	56	29.5	OTU	18	20.5	OTU	45	31.8
OTU	60	29.7	OTU	19	32.7	OTU	46	32.3
OTU	61	29.7	OTU	20	23.5	OTU	47	37.1
OTU	62	29.1	OTU	21	15.8	OTU	48	24.7
OTU	63	29.6	OTU	22	26.8	OTU	57	36.1
OTU	64	25.5	OTU	23	38.3	OTU	58	31.9
OTU	65	29.8	OTU	24	31.8	OTU	59	39.4
OTU	66	26.8	OTU	25	38.2	OTU	78	30.8
OTU	67	28.2	OTU	26	22.6	OTU	79	32.3
OTU	68	29.5	OTU	27	40.3*	OTU	80	22.8
OTU	69	29.4	OTU	28	33.4	OTU	81	23.8
OTU	70	28.4	OTU	29	34.7	OTU	82	22.5
OTU	71	30.0	OTU	30	39.2	OTU	83	17.3
OTU	72	29.5	OTU	31	23.1	OTU	84	24.3
OTU	73	30.0	OTU	32	37.5	OTU	85	31.0
OTU	74	29.5						
OTU	75	27.1						
OTU	76	30.0						
OTU	86	30.0						

*OTU BEYOND 5% LIMIT FOR NORMAL DISTRIBUTION

15.8-43.5 (Table 6). One OTU, *R.hookeri an.* (37), just exceeded the critical value as an outlier at $p<0.05$. The canonical variate analysis of these two groups gave excellent overall discrimination ($p<0.001$, Fig.4) with only one OTU (11) being considered misclassified. This OTU did not feature in the previous analysis as having an abnormally large D^2. The one OTU (37) that did feature in the previous analysis was orientated in the general direction of the alternate diploid centroid, but its 'outlier' status was such that it was placed closer to the polyploid centroid. The variables considered important in the standardised canonical function did not represent those with unusual distributions as shown by the two single group analyses. The larger multivariate measure of kurtosis for the non-diploid group is apparently reflected in the wider spread along the canonical axis, and the multivariate skewness shown in the diploid subset is mirrored by a skewed distribution on the canonical axis.

(iii) The final subdivision of the data set into those OTUs inhabiting riverbeds ($n=39$), and those that do not ($n=47$), gave z-scores for multivariate measures of skewness and kurtosis of 2.35, 0.47 and 0.12, 1.33 respectively. The riverbed subset had significant non-normality in 6 variables, 3 of these being floral counts. None of the OTUs within this group approached the critical ($p<0.05$) χ^2 value for the Mahalanobis distances of the OTU to the group centroid (Table 7). The non-riverbed subset showed extreme non-normality in 8 variables, and despite the indications of a more normal multivariate distribution, had a larger range of D^2s, 15.1-41.4 as against 18.0-36.1 for the riverbed subset (Table 7). The value of 41.4 just exceeding the critical value of 40.0 belongs again to *R.cinerea* (77). The two-group canonical variate analysis separated the two groups very well ($p<0.01$; Fig.5) with the classification in terms of minimum D^2s being a successful 91.9%. The 7 OTUs that were misclassified had shown no indication of being outliers from their respective groups in the previous single group analyses and the one OTU that did show up as being an outlier (77) was not misclassified but had the largest distance of any OTU to its 'correct' centroid.

DISCUSSION:

The assessment of the multivariate normality of an individual taxon within a taxonomic hierarchy is not recorded as having been attempted before, and the results from this analysis are encouraging as far as the numerical taxonomist is concerned. The full data set, although not strictly normal as defined by the multivariate measures of skewness and kurtosis, was sufficiently close

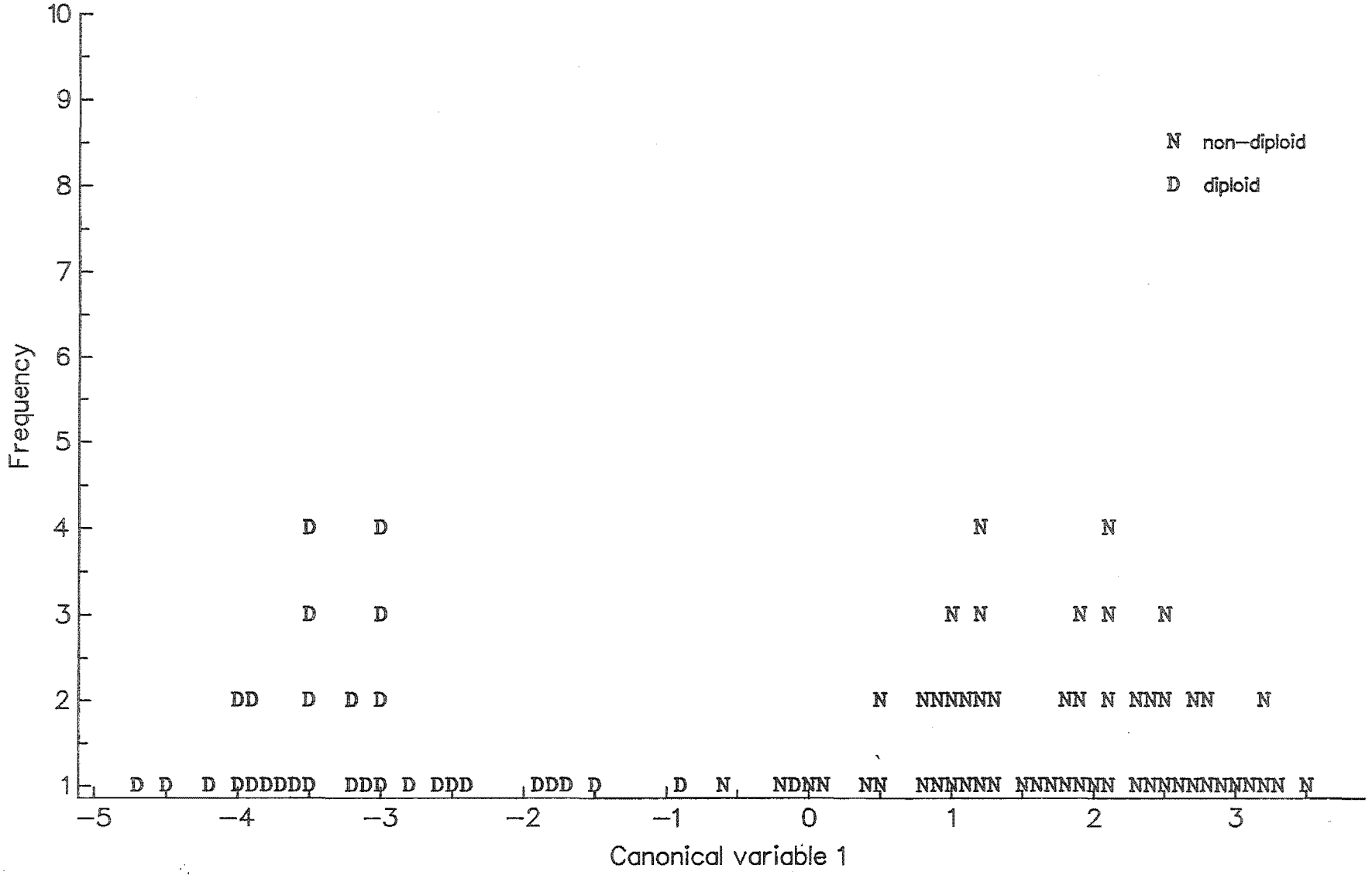

Figure 4: Plot of canonical variable 1 for diploid and non-diploid groups.

TABLE 7: MAHALANOBIS D^2 S FOR RIVERBED AND NON-RIVERBED GROUPS

RIVERBED			NON-RIVERBED					
OTU	9	25.8	OTU	1	30.9	OTU	42	18.8
OTU	10	32.1	OTU	2	25.1	OTU	43	27.1
OTU	12	32.3	OTU	3	26.6	OTU	44	27.0
OTU	13	27.5	OTU	4	27.5	OTU	49	39.5
OTU	16	18.7	OTU	5	36.2	OTU	50	35.7
OTU	18	20.6	OTU	6	26.6	OTU	54	18.5
OTU	19	36.1	OTU	7	25.5	OTU	55	28.1
OTU	20	18.0	OTU	8	21.5	OTU	56	27.2
OTU	21	27.5	OTU	11	26.2	OTU	57	36.6
OTU	23	29.7	OTU	14	19.3	OTU	58	35.8
OTU	24	33.7	OTU	15	34.0	OTU	59	36.8
OTU	25	28.2	OTU	17	30.2	OTU	74	30.7
OTU	27	30.3	OTU	22	34.5	OTU	75	31.2
OTU	28	30.4	OTU	26	30.2	OTU	76	22.2
OTU	29	33.9	OTU	30	37.2	OTU	77	41.4*
OTU	34	30.5	OTU	31	34.0	OTU	78	26.8
OTU	35	31.0	OTU	32	36.6	OTU	79	29.8
OTU	45	30.5	OTU	33	36.1	OTU	80	22.6
OTU	46	27.2	OTU	36	25.3	OTU	81	21.6
OTU	47	34.8	OTU	37	36.5	OTU	82	28.9
OTU	48	25.6	OTU	38	28.7	OTU	83	20.5
OTU	51	30.2	OTU	39	28.5	OTU	84	24.0
OTU	52	34.6	OTU	40	15.1	OTU	85	31.7
OTU	53	25.8	OTU	41	28.5			
OTU	60	28.1						
OTU	61	32.4						
OTU	62	33.7						
OTU	63	33.0						
OTU	64	25.2						
OTU	65	34.5						
OTU	66	32.6						
OTU	67	25.8						
OTU	68	21.4						
OTU	69	30.0						
OTU	70	23.3						
OTU	71	31.0						
OTU	72	27.4						
OTU	73	31.0						
OTU	86	26.7						

OTU BEYOND 5% LIMIT FOR NORMAL DISTRIBUTION

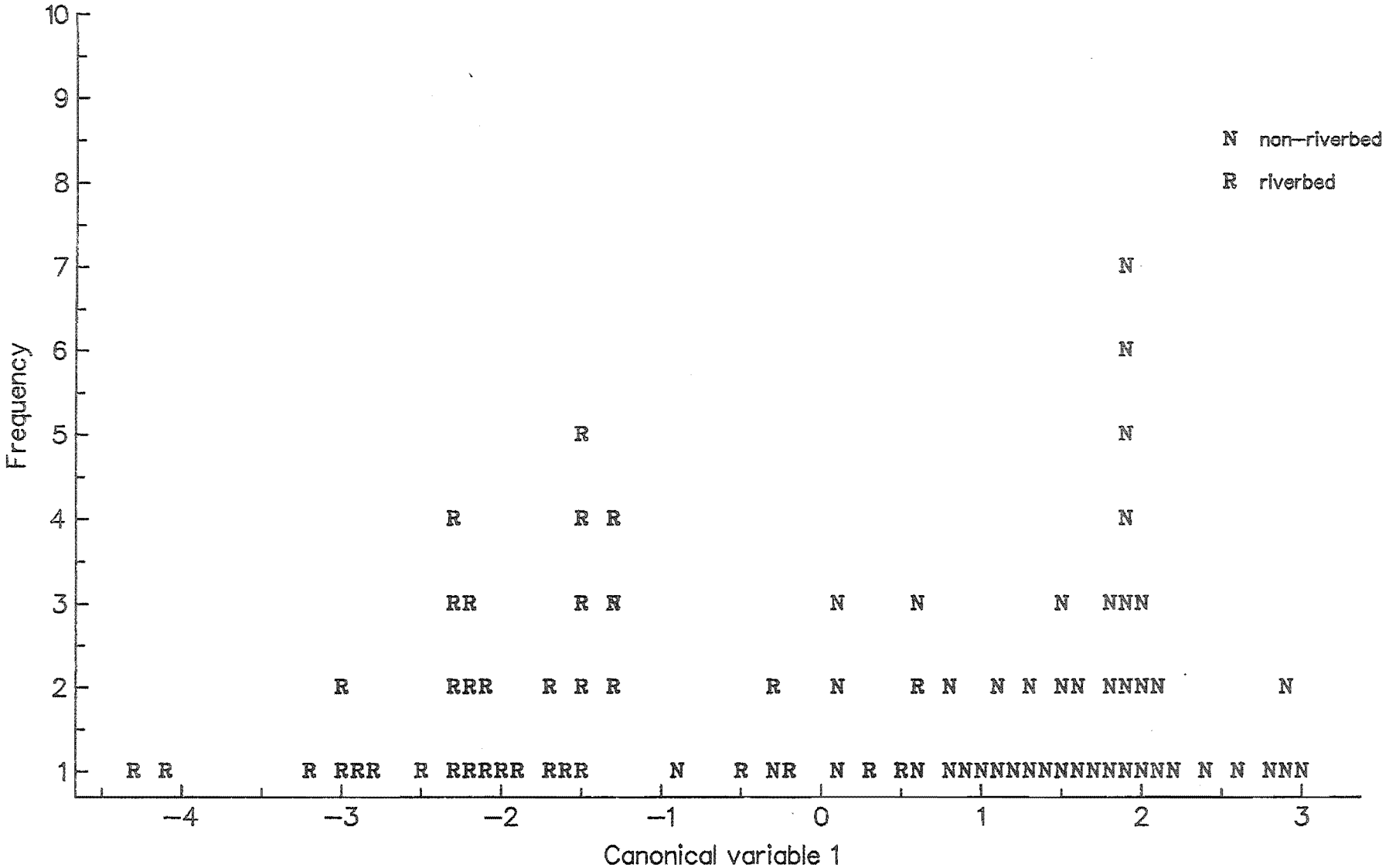

Figure 5: Plot of canonical variable 1 for riverbed and non–riverbed groups.

enough to justify sampling of the type used for this data, although some obvious guidelines for the sampling are implied.

The most important of these is that univariate normality *per se* will not lead to multivariate normality. The deletion of the non-normal variables significantly disrupted the multivariate distribution. There are two possible explanations for this:

(a) With a comparatively small number of OTUs the analyst should be wary of under-sampling variables when completing a data set. The addition of more variables, even highly correlated ones, has never been shown to harm the form of the multivariate clusters (Rohlf, 1967), but the lack of an adequate number does not easily lead to the creation of a regular multivariate distribution.

(b) The non-normal univariate distributions may well be indicative of subclusters within the taxon under study. Thus univariate non-normality may reflect diagnostic characters for the subclusters within the taxon and those subclusters themselves are an inherent by-product of variation within a multivariate normal distribution. The lack of the non-normal characters, deleted in the present exercise, may thus be creating a false, distorted picture of the multivariate distribution of the taxon.

The detection of OTU outliers via the Mahalanobis D^2s is clearly a useful by-product when measuring multivariate normality via the measures of skewness and kurtosis. The degree of influence that a few individual OTUs have on the multivariate distribution will obviously be dependent on the sample size. With the comparatively small sample in this example (statistically speaking, not in the context of numerical taxonomy), disruption caused by the six 'outliers' was significant but not extreme. The six OTUs involved are outliers from the subgenus in terms of the variables here measured and the rigid criterion placed on non-outliers, but their actual biological associations within the taxon do not support their exclusion from the subgenus, (Ward pers. comm. 1987). The exception to this is *R.cinerea* (77), which by any definition is distinct from the other species in the subgenus. The remaining five OTUs, however, may represent a further example of a problem common to many biostatistical studies, namely the influence of a general size factor on the relationships revealed by such studies. There is no easy solution to this (although some have been proposed, e.g. Humpheries et al., 1981; Somers, 1986) for the obvious reason that size in

general is a phenotypic and genotypic trait among all living organisms. A likely solution to the problem in this context would involve further sampling of the taxon to theoretically reveal a less clearly defined juncture between the very large OTUs and the remaining OTUs, and to thus highlight as outliers only those OTUs with a unique core of correlated characters, different from that of the hypothesised group being studied.

The cluster analysis both with and without the outlying OTUs revealed two important points in relation to the detection of outliers. The first is that obvious outliers are detectable with cluster analysis. The second is that a cluster analysis of the type used here, i.e., Gower's range coefficient and the UPGMAA clustering, while not revealing the presence of OTUs that vary considerably from the group as a whole, is strongly influenced by them. The pair-group method of clustering highlights the similarities between any two close OTUs at the expense of the overall associations between clusters of 2 or more OTUs. Thus the removal of the outliers led to significant repercussions in the portrayal of the general form of the species involved, and did not, as one would like, (ie when the sample size was very large), leave the general associations unchanged.

The deletion of the ratio variables led to a less normal distribution. This is the likely product of a reduced character number not adequately describing the subgenus. The implication is that these ratios are a useful addition to the character set and help to describe the overall pattern of the subgenus. To test accurately the hypothesis of a useful addition, one would need to sample randomly character sets of the same size as the set in question without the ratios ($p=37$), and then to compare the multivariate measures of skewness and kurtosis for these random sets (bootstrapping). However, my earlier results gave no indication that the ratios *per se* are causing any disruption to the form of the subgenus as revealed by the other characters.

The results from the pair-group discriminant analyses and the associated tests for multivariate normality offer some useful insights for practitioners of any discriminant analysis. The most important point arising is that of the relationship between two of the assumptions for discriminant analysis:

(i) individual group multivariate normality

(ii) any OTU in the analysis belongs to one of the hypothesised groups within the analysis.

If one first tests the overall group for multivariate normality, the second assumption here can be tested readily and then met by OTU exclusion. If this second assumption is not tested, outlying OTUs may well bias the position of the centroid of their hypothesised group without necessarily standing out sufficiently to be considered 'misclassified' in a discriminant analysis. Therefore, if overall multivariate normality has not been tested in the situations where there is group overlap, as revealed in the canonical variate plots, not only the 'dubious' classifications but also the OTUs showing large D^2s from their group centroids should be scrutinised. In this way new groups may be hypothesised that did not feature in an original analysis.

If the discriminant analysis indicates that a certain OTU belongs in one particular group, rather than in any other, the individual multivariate normality of that group can indicate whether the OTU in question is actually outside the scope of the analysis. Any OTUs that are initially of an unknown status should be 'jackknifed' into the analysis and its minimum Mahalanobis D^2 (OTU to centroid) can be approximately tested using the χ^2 statistic with p degrees of freedom. Outlier status *per se* should be considered of equal importance, to potentially slight multivariate deviations in the direction of another group, (misclassifications). Given that individual group normality is tested and met, then discriminant analysis addresses the question of whether the character set being sampled differentiates the groups as defined. If the character set is considered to be a good representation of the taxa then genuine group associations can be revealed by a principal coordinate analysis of the D^2 matrix. The above situation is one typically encountered by numerical taxonomists, and the clear definition of the individual groups, e.g. by a study of its multivariate normality, is important before any comparisons with other groups are made.

In a more general (non-taxonomic) situation where group description, and thus definition, is not a problem, (as in the preceding examples diploid/polyploid and riverbed dweller/not), the only question being considered is whether the characters sampled can differentiate the groups. If diagnostic characters or character combinations are being sought, then a stepwise discriminant analysis is appropriate. Although the testing of single group multivariate normality is clearly desirable, the frequently occurring problem, particularly for numerical taxonomists, is that of obtaining an adequate sample size to enable this test to be carried out.

The individual multivariate normality assessments reinforce the earlier contention relating the sample size to the degree of non-normality. This is evident in the skewness measures where the very large values were always associated with the smaller sample size. With multivariate examples, as with univariate equivalents, a certain minimum number of OTUs is required to assess normality. Although not specifically studied here, the impression obtained is that as p increases the requisite minimum sample size decreases. These results in general support the findings of Reyment (1971), but they are applicable on a different scale given the much larger number of variables.

The consistent correlation between a larger range of the Mahalanobis D^2s (OTU to centroid) and a more 'normal' distribution is in keeping with the ideas of a symmetrical non-clumped distribution for any normal single variable. In all instances this larger range was not associated with a large skewness measure, but is seen to be 'normal' variation about a centroid.

The association that occurred in the diploid/polyploid comparison between the distribution on the canonical variate axis and the multivariate measures of skewness and kurtosis may be a chance result. It relies primarily on the alignment of the principal axis of the individual group with the discriminating axis between the groups. Nevertheless, this situation is not necessarily unlikely to occur and it highlights the relationship between multivariate normality and discrimination.

The association between 'normal' within-group variables and effective discriminators is not a surprising one. It reinforces the testing of within-group univariate normality rather than overall normality, specifically if one is concerned to identify good discriminating characters. For it seems likely that these good discriminating characters are non-normal for the overall group but are in fact an important component in the multi-modal distribution of the characteristics of the taxon as a whole.

CONCLUSION:

A program has been written that will assess the multivariate measures of skewness and kurtosis for any data set, given that n is greater than p. This program has been tested on the *Raoulia* data set and has given results that indicate the practical significance of such measures to numerical taxonomists. The multivariate non-normality as indicated by these measures was found to be more a product of OTU outliers than of non-normal characters. The ratio characters were

shown to be relevant additional characters in terms of the overall normality of the data set. Indications from the removal of the non-normal characters and the ratio characters are that, in sampling a taxon within the framework of numerical taxonomy, the undersampling of variables is as great a source of difficulties as the omission of OTUs.

The procedure of investigating the multivariate normality of individual groups as a preliminary step before discriminant analysis is shown to be of practical use as well as being a statistical pre-requisite. The correct identification of single group outliers, whether they belong outside the postulated groups of a study, or whether they are really misclassified within the study, can be properly assessed only when the group is first studied independently and then simultaneously with the other groups.

The latent influence of 'moderate outliers' on a cluster analysis is shown to be considerable. This supports the use of more than one multivariate technique when elucidating the general structure of a multivariate data set.

CHAPTER 4

DISCRIMINANT ANALYSIS

"The business of a poet, said Imlac, is to examine, not the individual, but the species;... he does not number the streaks of the tulip, or describe the different shades in the verdure of the forest." SAMUEL JOHNSON

INTRODUCTION:

Since Fisher's (1936) exposition of the linear discriminant function and the subsequent developments of Mahalanobis (1936), Bartlett (1947), and Rao (1952), discriminant analysis has become a widely accepted tool for multivariate analysis. An elegant portrayal of all the relevant methodology may be found in Lachenbruch (1975). Different scientific fields in which the techniques of discriminant analysis have been utilised include, with examples, botany: Green (1974), Del Moral (1975), Collins et al. (1981), Hopper & Campbell (1977); zoology: Atchley (1974), Albrecht (1979), Seidel & Lucchino (1981); education: Porebski (1966), Anderson et al. (1969), Baggaley & Campbell (1967); geology: Saha & Rao (1971), Neilsen et al. (1973); medicine: Titterington et al. (1981), Truett et al. (1967); marketing: Gatty (1966); and astronomy: Nathanson (1971).

Despite the abundance of examples, many of which use accepted statistical packages, there remains some ambiguity in the literature with regard to aims and to the specific technique being used. The major problem relates to the definitions of such terms as 'Fisher's linear discriminant analysis', 'multi-group discriminant analysis', 'canonical variate analysis', 'stepwise discriminant analysis' and 'classification analysis'. Two papers (Kshirsagar& Arseven, 1975; and Samathanan, 1975) do in fact clarify any ambiguities that may exist by exploring the mathematical basis of the techniques involved. They clearly show the development of the classification function and its relationship to the original Fisher's linear discriminant function. They further show that the multi-group extension, canonical variate analysis, is essentially the non-trivial generalisation of Fisher's discriminant analysis. The above three techniques utilise the entire data set, excepting only the

deletion of non-significant canonical variables, but any one axis in a multi-group example is constructed to have a different emphasis, be it in a classification analysis or a canonical variate analysis. An axis from a classification analysis will maximise the discrimination of any single group from the remaining data points, while the sequential canonical variate axes successively portray the maximum ratio of between-group dispersion to the average within-group dispersion over all groups. It can be shown that when all canonical variables are retained, the classification of an individual based on a minimum Mahalanobis D^2 is equivalent to maximising its classification score over all groups (Kshirsagar& Arseven, 1975).

Jackknife classifications based on minimum Mahalanobis D^2s provide a useful tool, specifically for the numerical taxonomist, for assessing the true position of any OTU for which classification is undecided. They do this by portraying the relationships of a single OTU to the groups in the analysis, without using this same OTU to calculate the sample mean vectors (centroids) or the sample pooled within-group variance-covariance matrix. In so doing, the technique provides a further method for detecting outlying OTUs, (that is, those OTUs with relatively large jackknife D^2) and goes some way to overcoming the inherent bias in discriminant analysis to reproduce the hypothesised grouping structure.

The projection of points on to the first q canonical variables is equivalent to a representation of a principal co-ordinate analysis of the D^2 values (calculated using the pooled variance-covariance matrix) in q dimensions (Gower, 1966). This provides an accurate technique for viewing a multivariate distribution in fewer dimensions, while compensating for within-group inter-character correlations.

The method of stepwise discriminant analysis utilising different 'stepping' criteria develops a subset of variables that provide 'specific' maximum discrimination. The aim with stepwise discriminant analysis is not to reproduce a multivariate picture in a perceptible number of dimensions but rather to find a diagnostic function that differentiates all groups. If the majority of the groups are not colinear, this technique is frequently biased toward discriminating the most distinct groups. This method presupposes an exact knowledge of group classification.

Further problems arise from the use of discriminant analysis in the form of the statistical assumptions that need to be met if an accurate assessment of classifications is to be made. The

problem of multivariate normality for each group has already been addressed. Suffice to say that in most examples of numerical taxonomy this assessment is unlikely to be possible because of the small samples that often constitute the sub-groups in a classification. The influence of outliers has also been addressed and their influence on multivariate normality shown. However, outliers *per se* will also disrupt the individual group variance-covariance matrices, tending to increase the pooled variance-covariance estimates, and thus to reduce the magnitudes of the D^2 and consequently the precision of the analysis. A further repercussion of the heterogeneity of the variance-covariance matrices will be an indeterminable degree of bias in the probabilistic interpretation of classifications. The bias if it is of a consistent proportional type may tend to underestimate distances within small, closely related OTUs and to 'explode' the form of less integrated groups. Once again the small sample sizes typical of numerical taxonomy do not enable the testing of the homogeneity of the variance-covariance matrices by techniques such as Box's M (in Cooley and Lohnes, 1971) or Bartlett's (1947). However there is a technique, used by Campbell & Mahon (1974), which can be used to determine the effects of any unequal variance-covariance matrices. This involves the calculation of the first few principal co-ordinates of the D^2 matrix, using both the individual group variance-covariance matrices and the pooled variance-covariance matrix to calculate the D^2 matrix. By comparing the two plots, the influence of any heterogeneity can easily be seen and robust associations identified. A number of other papers have addressed the problem of the variance-covariance assumption (Melton, 1963; Krzanowski, 1977; Pimental, 1981; Marks & Dunn, 1974). It is not within the scope of this thesis to address this problem, but it may be said that the disproportionate sample sizes of the 'species' groups within the subgenus being studied may lead one to expect heterogeneity in the variance-covariance matrices and thus to view the absolute probabilities from the subsequent analyses with some caution.

The original theory of discriminant analysis was developed from the use of continuous variables and some contention exists regarding its application for data sets involving mixed variable types. The technique obviously makes use of the effects of an ordinal linear scale, thus making unordered multi-state characters inappropriate. The employment of discriminant analysis on binary data, with or without continuous variables is, however, feasible. The paper by Maxwell (1961) is frequently quoted as the justification for such use. Further papers by Krzanowski (1975, 1980), while advocating an alternative method, have generally supported this use. Given then that mixed

data sets of this form are appropriate to discriminant analysis, one can, bearing in mind the obvious character weighting and the creation of highly correlated characters, recode multi-state characters into a number of binary characters and employ them also. This latter data set type would be most appropriate when discrimination was the primary aim, rather than a lower dimensional representation of the multivariate distribution.

Discriminant analysis will be applied to this data set as a means of elucidating the multivariate associations between the 'species' groups, in the manner of a principal co-ordinate analysis. It is not being used in this context to specifically discriminate groups or sets of groups of the individual species.

METHOD:

BMDP (Dixon, 1987) program 7M was used to perform the discriminant analysis on this data set. The package provides the option of entering all variables, given that they pass a certain minimum tolerance (0.001), (where tolerance is one minus the multiple coefficient of determination of a single variable with all others) and the output can thus be viewed as a principal co-ordinate analysis of the Mahalanobis D^2s using the pooled within-groups variance-covariance matrix. The program also provides the alternative of jackknife or non-jackknife Mahalanobis D^2 on which to assess OTU misclassification. As with any discriminant analysis program, any variables that have no within-group variability cannot be entered. This enables the pooled variance-covariance matrix to be inverted.

The initial grouping structure will be that of Ward (1982) (Appendix A) with the single OTU species *R.cinerea* being omitted. This species is not entered not only because it has been shown in the previous study to be an outlier, but also because the bias generated by a single OTU group with no within group dispersion is likely to underestimate the pooled variance-covariance matrix and thus overestimate the D^2s, thereby entering a degree of uncertainty into the analysis. This leaves the number of groups at 13 and the number of OTUs at 85.

The variables used include all continuous characters together with the binary alternate characters. Given that the aim of this analysis is to portray the group structure and inter-group associations rather than to identify specific diagnostic characters and latent factors, given a satisfactory classification, the non-ordered multi-state characters have not been recoded into a

number of highly correlated binary characters. This gives a total of 74 characters available for analysis.

The pattern of the present analysis is one that provides for a progression of individual discriminant analyses, each incorporating changes that have been suggested by the results of the previous analyses. The criterion for a 'satisfactory' final classification can not be predefined. This should minimise mis-classifications and should not contain OTUs that have excessive minimum jackknife D^2s, given the general magnitude of the D^2s.

RESULTS:

Analysis 1: Groups as for Ward (1982)

Five variables (16, 30, 93, 95, 96) were excluded from this analysis as having no within-group variance. A further eight failed the tolerance criterion for inclusion (31, 41, 50, 53, 59, 64, 73, 75). The first two canonical variates accounted for 73% of the total dispersion. The total dispersion amounted to 1373.

From both the canonical plot (Fig.6) and the jackknife D^2s, (Table 8) *R.parkii* shows a close affiliation with *R.hookeri as.*. Three of the eight OTUs (3, 7, 8) are closer to the *R.hookeri as.* centroid than to the *R.parkii* centroid, and a further two OTUs (1, 6) have a minimum jackknife D^2 that places them beyond a reasonable relative distance from any centroid. These two latter OTUs were both revealed as general outliers in the single group multivariate normal study. *R.hookeri as.* itself has no mis-classified OTUs, but all the OTUs in this species show a strong similarity to those of *R.parkii* (Table 8).

R.australis shows a strong affiliation with the *R.hookeri h.* species (Fig.6), with four of the OTUs (9, 15, 17, 85) having a smaller jackknife D^2 to this centroid than to the *R.australis* centroid. Of the 19 *R.hookeri h.* OTUs, three (24, 27, 31) are closer to the *R.australis* centroid than to the *R.hookeri h.* centroid and a further two (26, 30) show no clear affiliation with any species group. One of these OTU 30 was shown to be an outlier in the previous multivariate normal study. The remaining 14 OTUs display a close similarity to *R.australis.*

Five of the six OTUs of *R.hookeri an.* (40, 41, 42, 43, 44) are shown to be very close to the two-OTU group *R.beauverdii* (38, 39). Two of these five (40, 41) are closer to the centroid of

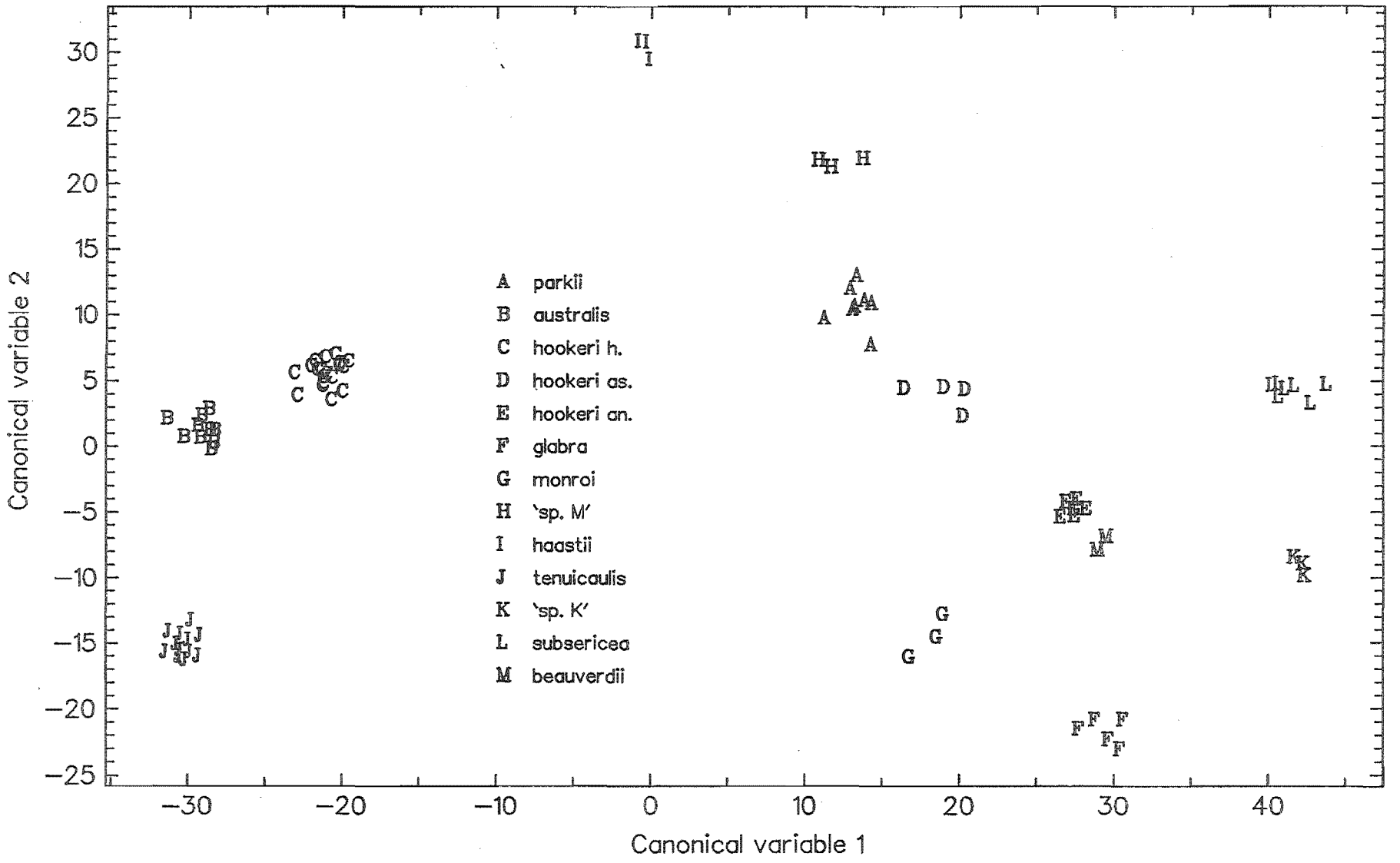

Figure 6: Plot of canonical variable 1 against canonical variable 2 for 13 species groups.

TABLE 8: MAHALANOBIS D^2S FOR THE INITIAL GROUPS

		D^2 TO OWN CENTROID	D^2 TO CLOSEST SPECIES
1)*R.parkii*			
OTU	1	xxxx.x	xxxx.x
OTU	2	932.0	1467.9[8]
OTU	3	796.7[4]	753.3[6]
OTU	4	510.9	1298.2[6]
OTU	5	505.5	690.1[6]
OTU	6	1981.2	2137.9[6]
OTU	7	704.9[4]	534.4[6]
OTU	8	724.7[4]	455.8[6]
2)*R.australis*			
OTU	9	940.4[4]	667.6[3]
OTU	10	924.5	1683.9[3]
OTU	11	528.1	881.4[3]
OTU	12	715.8	734.6[3]
OTU	13	563.1	1198.0[3]
OTU	14	348.6	753.0[3]
OTU	15	808.8[4]	790.8[3]
OTU	16	306.1	433.5[3]
OTU	17	654.1[4]	265.1[3]
OTU	85	400.4[4]	351.8[3]
3)*R.hookeri h.*			
OTU	18	226.4	270.2[2]
OTU	19	392.1	392.9[2]
OTU	20	558.9	843.4[2]
OTU	21	368.6	790.7[2]
OTU	22	282.4	550.8[2]
OTU	23	598.4	613.1[2]
OTU	24	796.2[4]	790.8[2]
OTU	25	742.6	953.7[2]
OTU	26	6473.2[4]	1344.6[1]
OTU	27	534.5[4]	471.42
OTU	28	1142.4	1344.5[1]
OTU	29	964.6	1536.6[2]
OTU	30	6472.8[4]	6395.7[1]
OTU	31	334.4[4]	297.0[2]
OTU	32	1011.5	1058.7[2]
OTU	45	483.0	695.4[2]
OTU	46	527.0	1034.6[2]
OTU	47	1018.9	1809.7[2]
OTU	48	159.4	358.6[2]
4)*R.hookeri as.*			
OTU	33	165.7	432.2[1]
OTU	34	1271.0	1409.3[1]
OTU	35	370.1	537.8[1]
OTU	36	223.5	379.7[1]
5)*R.hookeri an.*			
OTU	37	xxx.x	xxx.x
OTU	40	201.1[4]	198.1[13]
OTU	41	349.7[4]	241.9[13]
OTU	42	126.5	218.8[13]
OTU	43	124.0	125.1[13]
OTU	44	293.3	370.1[13]
6)*R.glabra*			
OTU	49	896.0	2063.3[11]
OTU	50	922.7	1478.0[11]
OTU	51	1210.2	2311.7[12]
OTU	52	2023.7[4]	1214.6[11]
OTU	53	509.1	1660.3[8]
7)*R.monroi*			
OTU	54	690.1	1763.2[6]
OTU	55	916.5	1038.7[5]
OTU	56	208.4	908.1[5]
8)*R.'sp.M'*			
OTU	57	681.1	1441.2[1]
OTU	58	335.5	1535.4[1]
OTU	59	803.6	2807.6[11]
9)*R.haastii*			
OTU	60	976.4	2857.1[8]
OTU	61	1276.7	1965.9[1]
OTU	62	1468.5	2175.2[1]
10)*R.tenuicaulis*			
OTU	63	1915.3	2170.6[2]
OTU	64	349.8	763.5[2]
OTU	65	1383.4	1931.1[2]
OTU	66	616.3	2377.5[2]
OTU	67	348.1	688.6[2]
OTU	68	316.6	1556.9[2]
OTU	69	366.9	844.1[2]
OTU	70	152.0	948.8[2]
OTU	71	351.0	1218.3[2]
OTU	72	1068.6	2499.1
OTU	73	839.1	1273.9[2]
OTU	86	1068.6	1305.6[2]
11)*R.'sp.K'*			
OTU	74	4238.7[4]	1440.8[7]
OTU	75	4238.7	4418.0[12]
OTU	76	xxxx.x	xxxx.x
12)*R.subsericea*			
OTU	78	999.7	1076.611
OTU	79	637.3	1907.5[8]
OTU	80	475.1	1051.7[11]
OTU	81	175.8	915.8[11]
OTU	82	288.6	850.8[11]
OTU	83	351.6	830.4[11]
OTU	84	428.4	1230.4[6]
13)*R.beauverdii*			
OTU	38	635.9[4]	311.2[5]
OTU	39	635.9[4]	198.5[5]

[1] *R.parkii*
[2] *R.australis*
[3] *R.hookeri h.*
[4] *R.hookeri as.*
[5] *R.hookeri an.*
[6] *R.glabra*
[7] *R.monroi*
[8] *R.'sp.M'*
[11] *R.'sp.K'*
[12] *R.subsericea*
[13] *R.beauverdii*

[4] OTU MISCLASSIFIED

that group than to the centroid of the *R.hookeri an.* species. OTU 37 shows a lack of association with any of the species groups, a fact that was revealed in the previous chapter. The two OTUs in the *R.beauverdii* group (38, 39) are both closer to the *R.hookeri an.* centroid on the basis of their jackknife D^2s.

Four of the five OTUs of *R.glabra* (49, 50, 51, 53) show this species to be a tightly self contained unit. The only exception is OTU 52 which lies closer to the hypothesised *R.'sp.K'* than to *R.glabra*.

R.monroi, *R.'sp.M'*, and *R.subsericea* are all shown to be concentrated individual groups adequately distinct from any other species (Fig.6).

R.haasti and *R.tenuicaulis* appear to be on the periphery of the subgenus as a whole. For this reason they are both sufficiently distinct from any of the other species in the subgenus, but they lack the unity of the three species above (Table 8). None of the OTUs are misclassified, but some of the jackknife D^2s to the group centroids are larger than would be acceptable for the less well differentiated species. OTUs 62 and 63 are clear examples of this.

R.'sp.K' is shown to be an apparently random association of OTUs (Table 8). Two of the three (74, 76) are close to *R.monroi*, but none of the three have a sufficiently small jackknife D^2 to any centroid to indicate a convincing association.

Analysis 2: Re-analysis with deletions suggested above

Five OTUs 1,6,26,30 and 37 and *R.'sp.K'* were removed from the data set and the analysis rerun. Seven variables were excluded from this analysis as having no within-group variance. These were: 16,30,83,86,93,95 and 96. A further thirteen failed the tolerance criterion for inclusion. These were: 38,39,45,46,50,53,55,57,59,60,63,64,68. The first two canonical variates accounted for 67% of the total dispersion. this dispersion amounted to 812.

The removal of two OTUs from *R.parkii*, now leaves the group in disarray: four OTUs (3, 5, 7, 8) are now very close to the *R.hookeri as.* centroid and two (2, 4) show a strong association with *R.hookeri an.*, (Table 9). The general association with *R.hookeri as.* is still strongly evident,

TABLE 9: MAHALANOBIS D^2'S FOR THE INITIAL GROUPS WITH DELETIONS

		D^2 TO OWN CENTROID	D^2 TO CLOSEST SPECIES
1)R.parkii			
OTU	2	1191.3*	574.4[5]
OTU	3	169.1	280.5[6]
OTU	4	702.0*	700.0[5]
OTU	5	449.1*	419.1[4]
OTU	7	444.7*	397.7[6]
OTU	8	682.6*	474.3[6]
2)R.australis			
OTU	9	595.3	695.7[3]
OTU	10	594.1	1420.4[3]
OTU	11	687.4	1106.3[5]
OTU	12	619.9*	427.2[13]
OTU	13	308.1	754.5[3]
OTU	14	416.5	445.5[13]
OTU	15	933.7*	794.7[10]
OTU	16	257.8	284.0[3]
OTU	17	781.1*	411.1[3]
OTU	85	277.6	369.6[3]
3)R.hookeri h.			
OTU	18	136.7	202.3[2]
OTU	19	341.5	585.4[2]
OTU	20	429.0	649.2[2]
OTU	21	274.4	508.4[5]
OTU	22	222.3	391.5[2]
OTU	23	228.2	412.2[5]
OTU	24	312.2	565.6[5]
OTU	25	401.8	644.2[5]
OTU	27	514.1*	401.4[2]
OTU	28	595.9	641.5[2]
OTU	29	1134.2	1263.8[2]
OTU	31	372.9	527.9[2]
OTU	32	1054.5	1095.0[2]
OTU	45	315.2	402.0[13]
OTU	46	896.0*	468.8[5]
OTU	47	606.0	1180.5[5]
OTU	48	538.0	1187.1[2]
4)R.hookeri as.			
OTU	33	181.7	320.4[1]
OTU	34	1322.2	1618.3[10]
OTU	35	314.4	509.3[1]
OTU	36	383.9*	223.8[1]
5)R.hookeri an.			
OTU	40	194.9*	183.8[13]
OTU	41	154.2*	137.4[13]
OTU	42	173.4	348.7[13]
OTU	43	117.3	117.5[13]
OTU	44	494.5	499.8[13]
6)R.glabra			
OTU	49	790.2	2056.8[12]
OTU	50	1288.3	1739.2[4]
OTU	51	1301.3	2834.6[10]
OTU	52	1608.5*	1406.5[1]
OTU	53	579.2	1804.0[1]
7)R.monroi			
OTU	54	488.0	1333.4[2]
OTU	55	1106.9*	826.4[3]
OTU	56	259.2	963.6[1]
8)R. 'sp.M'			
OTU	57	815.7	2229.3[12]
OTU	58	389.9	2215.5[5]
OTU	59	455.7	1841.6[5]
9)R.haastii			
OTU	60	1386.8	3376.9[10]
OTU	61	2865.7*	1640.4[5]
OTU	62	1175.4	2259.3[1]
10)R.tenuicaulis			
OTU	63	1400.9	1511.8[2]
OTU	64	588.5	838.4[3]
OTU	65	923.6	1563.3[2]
OTU	66	689.1	2427.3[6]
OTU	67	386.4	744.5[3]
OTU	68	338.7	1452.8[2]
OTU	69	332.1	779.6[2]
OTU	70	106.3	1008.3[2]
OTU	71	326.2	1182.0[3]
OTU	72	746.1	2121.4[3]
OTU	73	322.7	837.9[3]
OTU	86	746.1	1064.5[2]
12)R.subsericea			
OTU	78	664.7	2390.5[4]
OTU	79	492.7	1582.7[1]
OTU	80	418.8	1139.9[4]
OTU	81	196.7	1061.0[1]
OTU	82	398.4	1174.6[4]
OTU	83	267.3	1854.8[4]
OTU	84	221.7	1166.1[1]
13)R.beauverdii			
OTU	38	911.9*	197.3[5]
OTU	39	911.8*	379.9[1]

[1] *R.parkii*
[2] *R.australis*
[3] *R.hookeri h.*
[4] *R.hookeri as.*
[5] *R.hookeri an.*
[6] *R.glabra*
[10] *R.tenuicaulis*
[12] *R.subsericea*
[13] *R.beauverdii*

*OTU MISCLASSIFIED

(Fig.7; Table 9) with one OTU (36) in this group closer to the *R.parkii* centroid than it is to its own *R.hookeri as.* centroid.

R.australis still shows a strong general affiliation with *R.hookeri h.*, (Fig.7; Table 9) with one (17) of the ten OTUs closer to this centroid than to the *R.australis* centroid. A further two OTUs (12, 15) are now considered misclassified, with 12 indicating the now closer association of *R.australis* with both *R.beauverdii* and *R.hookeri an.* and 15 indicating the similarity of this cluster to *R.tenuicaulis*. One OTU (27) from *R.hookeri h.* is closer to the *R.australis* centroid than to *R.hookeri h.*, and one further OTU (46) is considered misclassified, representing the close association of *R.australis* and *R.hookeri h.* with *R.hookeri an.* and *R.beauverdii*.

R.hookeri an. and *R.beauverdii* are still seen to be very similar (Fig.7; Table 9), with two of the *R.hookeri an.* OTUs (40, 41) and the two members of *R.beauverdii* (38, 39) showing a closer association with their respectively opposite centroids. One of these OTUs (39) from *R.beauverdii* is now closer to the *R.parkii* centroid than to its own centroid.

Within species *R.glabra*, OTU 52 still shows a dissimilarity to the other OTUs in the species. It now shows greatest affinity with the *R.parkii* group.

R.'sp.M', R.subsericea and *R.tenuicaulis* show little change as a result of the deletions made.

Within *R.monroi* and *R.haasti* two OTUs are now considered misclassified. OTU 55 is closer to *R.hookeri h.* indicating the close association now seen between *R.glabra* and *R.australis* and *R. hookeri h.* and OTU 61 is closer to the *R.hookeri an.* centroid.

Analysis 3: Re-analysis with suggested amalgamations

As strong pair associations have been shown between *R.parkii*–*R.hookeri as.*, *R.australis*–*R.hookeri h.* and *R.hookeri an.*–*R.beauverdii* these pairs were each combined and with *R.'sp.K'* still excluded the analysis was rerun. Six variables were excluded from this analysis as having no within-group variance. These were: 16,30,83,93,95 and 96. A further fourteen failed the tolerance criterion for inclusion. These were 39,41,43,44,50,55,57,59,61,63,64,68,70,and 77. The amount of dispersion accounted for by the first two canonical variate axes was 68 % the total dispersion was 484.

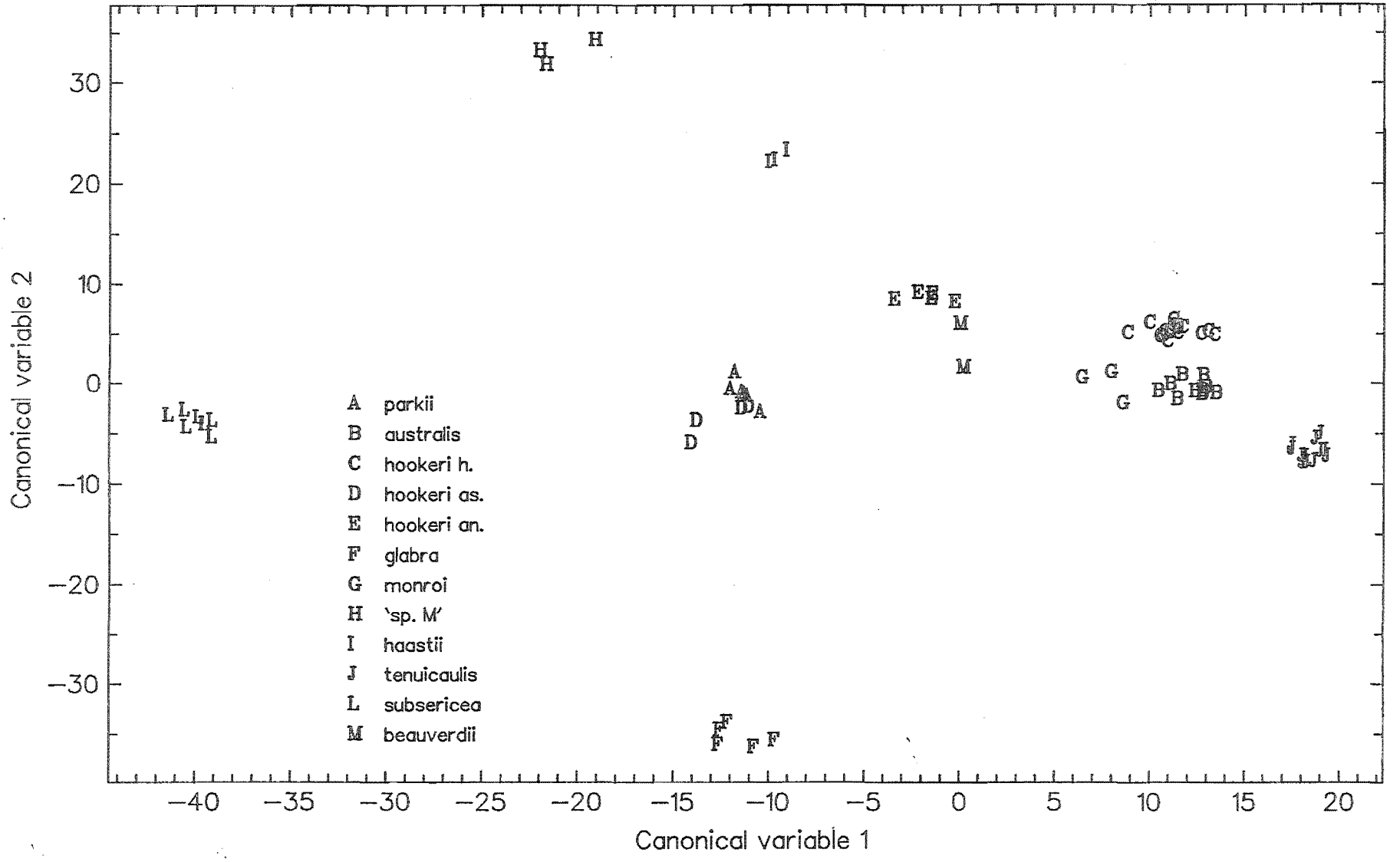

Figure 7: Plot of canonical variable 1 against canonical variable 2 for 12 species groups.

The newly-formed *R.parkii–R.hookeri as.* complex shows considerable unity (Fig.8), with only two OTUs considered misclassified, (Table 10). OTU 1 still shows its outlier status, and OTU 2 typifies the groups general similarity with the *R.hookeri an.–R.beauverdii* complex.

The *R.australis–R.hookeri h.* group also shows a strong within-group unity. Three of the OTUs (26, 30, 46) are now closer to the *R.hookeri an.–R.beauverdii* centroid on the basis of the jackknife D^2s, but one of these (30) has a minimum D^2 such that it fully justifies its earlier deletion. The other two exemplify the general association between this new group and the *R.hookeri an.–R.beauverdii* complex.

The remaining species groups, *R.hookeri an.–R.beauverdii*, *R.glabra*, *R.monroi*, *R.'sp.M'*, *R.haasti*, *R.tenuicaulis* and *R.subsericea*, all now show much 'better' structure in terms of within group homogeneity and discrimination from the other species. The one exception to this is the continued 'random' association of one *R.glabra* (OTU 52) with any species but *R.glabra*; on this occasion it is affiliated to *R.subsericea*.

Analysis 4: Both deleting and combining as suggested by the first analysis

The changes implemented in the previous two analyses were combined and the analysis rerun. Thus n=80 and g=9. Seven variables were excluded from this analysis as having no within-group variance. These were 16,30,83,86,93,95 and 96. A further twenty failed the tolerance criterion for inclusion. These were 28,34,35,37,38,39,41,43,45,53,54,55,59,60,63,64,68,70, 73, and 75. The amount of dispersion accounted for by the first two canonical variate axes was 67% the total dispersion being 382.

The deletion of the five OTUs has 'corrected' two previous misclassifications and added a further two, (Table 11). OTUs 46 and 52 are now classified into their hypothesised species. OTU 2 still shows a closer affiliation with the *R.hookeri an.–R.beauverdii* complex than with the *R.parkii–R.hookeri as.* complex. OTU 38 now has a smaller jackknife D^2 to the *R.australis–R.hookeri h.* centroid than to the *R.hookeri an.–R.beauverdii* complex, and OTU 44 is very marginally closer to the *R.parkii–R.hookeri as.* complex than to its own centroid. The structure of this final grouping arrangement is seen to be superior to the previous three, (Fig.9; Table 11). In

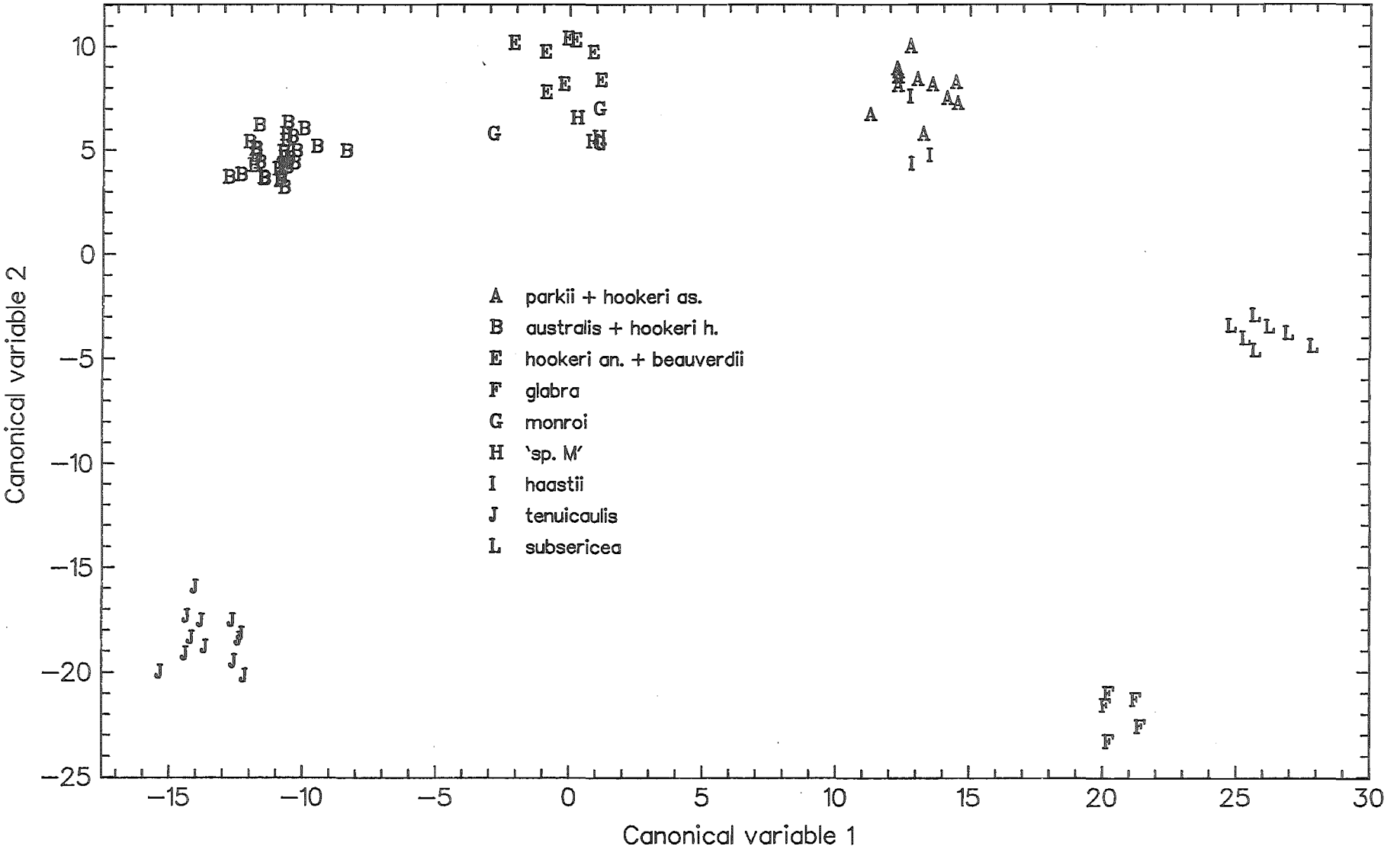

Figure 8: Plot of canonical variable 1 against canonical variable 2 for 9 species groups.

TABLE 10: MAHALANOBIS D^2'S FOR INITIAL GROUPS WITH AMALGAMATIONS

		D^2 TO OWN CENTROID	D^2 TO CLOSEST SPECIES
1)R.parkii + R.hookeri as.			
OTU	1	4398.8[4]	4359.9[6]
OTU	2	630.3[4]	370.7[5]
OTU	3	190.7	413.6[5]
OTU	4	284.5	460.3[5]
OTU	5	238.4	417.1[5]
OTU	6	506.4	841.0[12]
OTU	7	262.0	439.5[12]
OTU	8	258.8	364.1[5]
OTU	33	132.7	330.4[5]
OTU	34	815.1	1039.9[12]
OTU	35	219.4	451.7[5]
OTU	36	81.3	382.7[12]
2)R.australis + R.hookeri h.			
OTU	9	449.9	681.5[5]
OTU	10	285.1	498.6[5]
OTU	11	367.4	658.7[5]
OTU	12	314.7	455.3[5]
OTU	13	152.1	453.5[5]
OTU	14	256.2	467.2[5]
OTU	15	367.7	668.9[5]
OTU	16	153.7	442.7[5]
OTU	17	221.0	554.1[5]
OTU	18	71.6	344.8[5]
OTU	19	298.5	516.6[5]
OTU	20	231.0	446.3[5]
OTU	21	328.8	580.4[5]
OTU	22	243.8	668.6[5]
OTU	23	98.8	314.9[5]
OTU	24	151.1	367.7[5]
OTU	25	397.2	663.2[5]
OTU	26	1657.7[4]	220.8[5]
OTU	27	233.8	519.9[5]
OTU	28	398.6	574.1[5]
OTU	29	282.6	500.5[5]
OTU	30	1657.7[4]	1427.2[5]
OTU	31	230.6	593.1[5]
OTU	32	273.4	507.2[5]
OTU	45	125.5	422.9[5]
OTU	46	327.3[4]	318.8[5]
OTU	47	449.7	726.8[10]
OTU	48	157.7	395.5[5]
OTU	85	269.7	637.0[5]
5)R.hookeri an. + R.beauverdii			
OTU	37	102.2	274.5[1]
OTU	38	106.6	324.6[2]
OTU	39	91.1	299.9[1]
OTU	40	144.2	343.2[2]
OTU	41	199.7	433.3[1]
OTU	42	131.2	251.4[2]
OTU	43	66.8	269.3[2]
OTU	44	411.8	586.0[1]
6)R.glabra			
OTU	49	346.9	963.5[12]
OTU	50	325.0	791.7[12]
OTU	51	1001.6	2013.9[10]
OTU	52	1151.7[4]	608.6[12]
OTU	53	355.1	809.9[12]
7)R.monroi			
OTU	54	301.3	904.0[1]
OTU	55	506.4	903.1[2]
OTU	56	114.3	805.5[1]
8)R.'sp.M'			
OTU	57	484.6	844.2[5]
OTU	58	156.4	975.9[5]
OTU	59	310.5	1010.0[5]
9)R.haastii			
OTU	60	586.5	2165.4[12]
OTU	61	1038.4	1484.8[1]
OTU	62	686.3	1728.3[1]
10)R.tenuicaulis			
OTU	63	370.2	850.8[2]
OTU	64	292.7	825.2[2]
OTU	65	587.9	1167.1[2]
OTU	66	256.9	1272.3[2]
OTU	67	173.5	563.4[2]
OTU	68	264.5	1113.0[2]
OTU	69	245.4	703.6[2]
OTU	70	115.7	797.7[2]
OTU	71	286.9	1066.5[2]
OTU	72	649.3	1543.2[6]
OTU	73	502.5	811.9[2]
OTU	86	649.3	932.6[2]
12)R.subsericea			
OTU	78	538.6	1520.4[6]
OTU	79	137.6	543.7[1]
OTU	80	251.4	443.8[1]
OTU	81	115.6	500.6[1]
OTU	82	127.4	418.4[1]
OTU	83	131.6	611.4[1]
OTU	84	157.3	467.4[1]

[1] *R.parkii + R.hookeri as.*
[2] *R.australis + R.hookeri h.*
[5] *R.hookeri an. + R.beauverdii*
[6] *R.glabra*
[10] *R.tenuicaulis*
[12] *R.subsericea*

[4] OTU MISCLASSIFIED

TABLE 11: MAHALANOBIS D^2S FOR INITIAL GROUPS WITH AMALGAMATIONS AND DELETIONS

		D^2 TO OWN CENTROID	D^2 TO CLOSEST SPECIES
1)R.parkii + R.hookeri as.			
OTU	2	295.6[*]	239.4[5]
OTU	3	132.4	337.8[5]
OTU	4	222.0	286.9[5]
OTU	5	190.9	316.8[5]
OTU	7	110.8	244.1[5]
OTU	8	205.7	351.3[5]
OTU	33	114.3	254.3[5]
OTU	34	475.8	683.6[1 2]
OTU	35	192.6	202.8[5]
OTU	36	77.0	245.7[1 2]
2)R.australis + R.hookeri h.			
OTU	9	406.7	473.6[5]
OTU	10	239.2	458.3[5]
OTU	11	230.7	285.3[5]
OTU	12	228.3	314.9[5]
OTU	13	196.2	275.7[5]
OTU	14	195.9	283.5[5]
OTU	15	409.3	490.9[5]
OTU	16	123.3	150.5[5]
OTU	17	154.8	249.5[5]
OTU	18	59.7	160.0[5]
OTU	19	263.6	412.4[5]
OTU	20	195.8	233.4[5]
OTU	21	119.7	202.5[5]
OTU	22	146.8	256.6[5]
OTU	23	87.5	129.1[5]
OTU	24	160.8	172.9[5]
OTU	25	263.0	394.6[5]
OTU	27	262.2	439.4[5]
OTU	28	226.4	276.1[5]
OTU	29	299.9	354.0[5]
OTU	31	136.7	209.8[5]
OTU	32	203.9	246.9[5]
OTU	45	122.0	187.2[5]
OTU	46	231.1	269.3[5]
OTU	47	360.5	618.5[5]
OTU	48	124.3	163.9[5]
OTU	85	161.3	256.9[5]
5)R.hookeri an. + R.beauverdii			
OTU	38	202.0[*]	111.7[2]
OTU	39	122.5	166.9[1]
OTU	40	126.6	131.5[2]
OTU	41	78.8	187.6[1]
OTU	42	109.9	147.4[2]
OTU	43	47.1	126.8[2]
OTU	44	351.6[*]	349.9[1]
6)R.glabra			
OTU	49	380.2	1031.6[1 2]
OTU	50	462.0	934.1[1 2]
OTU	51	505.5	1690.8[1 2]
OTU	52	565.6	735.3[1 2]
OTU	53	386.4	953.4[1 2]
7)R.monroi			
OTU	54	216.2	628.5[1]
OTU	55	200.2	614.4[2]
OTU	56	96.1	564.5[1]
8)R.'sp.M'			
OTU	57	366.0	749.1[5]
OTU	58	145.6	708.5[5]
OTU	59	182.9	691.5[5]
9)R.haastii			
OTU	60	179.8	1520.7[5]
OTU	61	181.9	1319.3[5]
OTU	62	682.2	1394.2[8]
10)R.tenuicaulis			
OTU	63	256.5	688.8[2]
OTU	64	176.7	601.4[2]
OTU	65	501.2	860.0[2]
OTU	66	293.3	1302.7[8]
OTU	67	92.1	518.4[2]
OTU	68	146.3	754.1[2]
OTU	69	216.4	568.8[2]
OTU	70	53.6	720.2[2]
OTU	71	233.8	895.5[2]
OTU	72	446.7	1203.7[8]
OTU	73	199.8	598.2[2]
OTU	86	446.7	735.8[2]
12)R.subsericea			
OTU	78	448.4	1222.2[8]
OTU	79	149.1	342.2[1]
OTU	80	275.3	288.5[1]
OTU	81	81.5	458.1[1]
OTU	82	110.9	366.6[1]
OTU	83	111.6	346.6[1]
OTU	84	108.2	328.9[1]

[1] *R.parkii + R.hookeri as.*
[2] *R.australis + R.hookeri h.*
[5] *R.hookeri an. + R.beauverdii*
[8] *R.'sp.M'*
[1 2] *R.subsericea*

[*] OTU MISCLASSIFIED

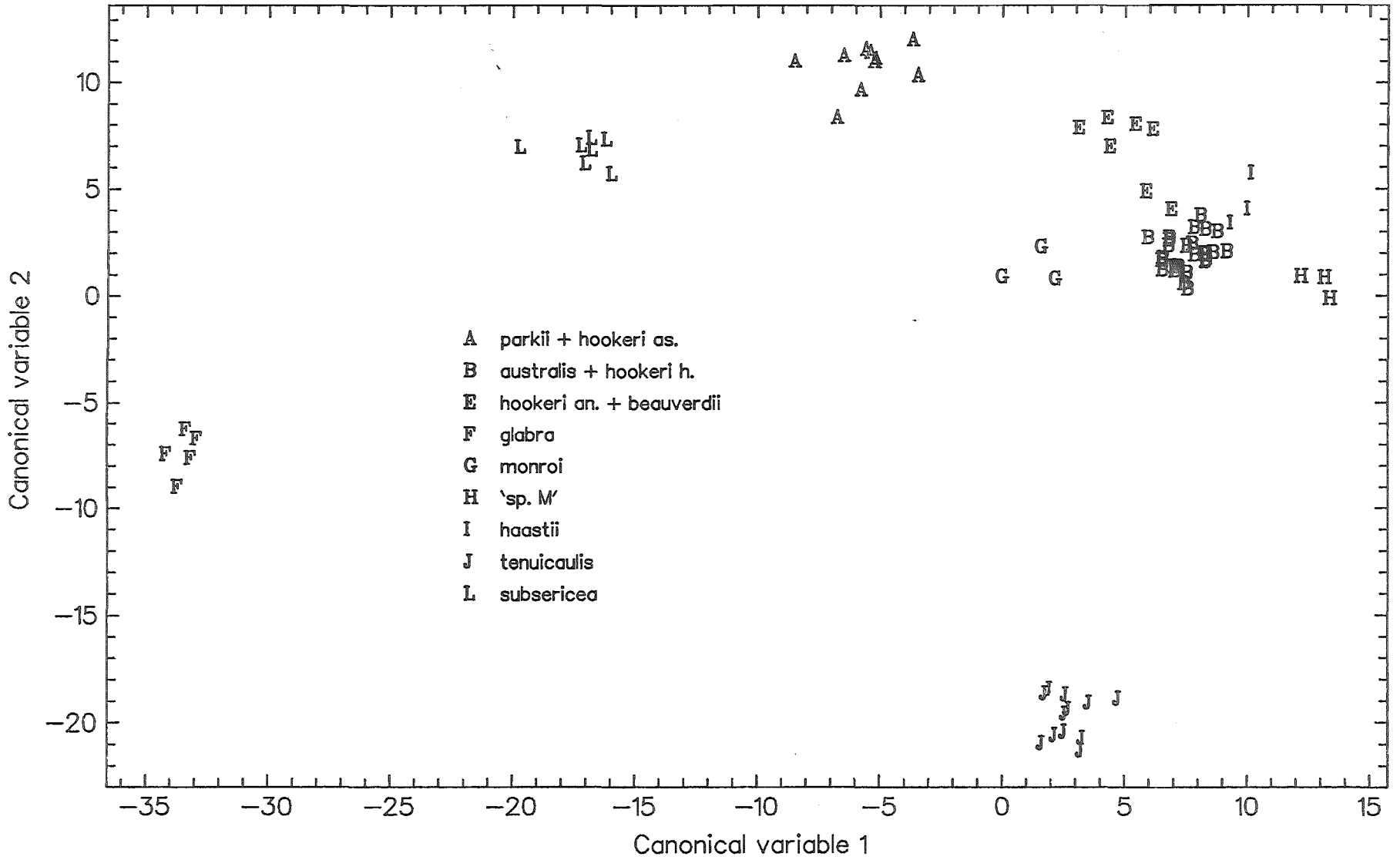

Figure 9: Plot of canonical variable 1 against canonical variable 2 for 9 species groups.

this analysis, as in the preceeding three, all non-jackknife classifications based on the minimum D^2 placed all OTUs in their 'correct' groups.

DISCUSSION:

The purpose of the preceding analyses has been to examine the validity of the hypothesised 'species' within the *Raoulia* subg. *Raoulia* group as defined by Ward (1982). In the process of altering the species compositions to form a validated classification the influence of a number of factors on an analysis of this form can be elucidated.

It is clear from the first analysis (Fig.6, Table 8), that within the subgenus the species as defined, while showing generally clear separation (Fig.6), are not homogeneous in terms of within group dispersion or in terms of minimum between group distances (Table 8). Five OTUs (1, 6, 26, 30, 37) stand out as outliers from the subgenus in general using the criterion of a minimum jackknife D^2, and four of these (1, 6, 30, 37) were similarly identified in the multivariate normal study. The blithe acceptance of any jackknife-judged misclassifications as the primary source of concern, when sample sizes are as small and as variable as they are here, would be unjustified. However, the relatively small number indicates some considerable robustness in most of the groupings. The completely successful non-jackknife classifications support both of these contentions. Generally the species *R.monroi*, *R.'sp.M'*, *R.haasti*, *R.glabra*, *R.tenuicaulis*, and *R.subsericea* appear to present no real problems. There is some disparity in some of these species in terms of within-group dispersion, but they are all clearly delimited from all the other species in the subgenus. The species *R.hookeri h.*, *R.hookeri as.*, *R.hookeri an.*, *R.australis*, *R.beauverdii* and *R.parkii*, however, do present some problems. Their inter-relationships indicate that they are much less well defined entities, and some likely recombinations are suggested by the first and second analyses. The three OTUs of *R.'sp.K'* show no clear association between themselves, while their relationships with the other species are inconsistent.

The second analysis (Fig.7; Table 9), whereby the outlying OTUs and *R.'sp.K'* are omitted, gives a clear indication of the impact of decisions to omit OTUs on all apparent associations of the OTUs. The strong inter-specific associations (*R.australis* & *R.hookeri h.*, *R.parkii* & *R.hookeri as.* and *R.beauverdii* & *R.hookeri an.*) are still evident, but there is some general disruption of the species centroids involved, notably that of *R.parkii*. There can be little doubt that this second

analysis is a more accurate picture of the true relationships, given this data set. It would appear, nevertheless, that when small sample sizes are involved even gross outliers, (OTUs 1, 6) detected by the methods used here, can provide some stability to certain individual centroid estimates. The course of the OTU deletions has increased the associations between the three pairs of species seen in the first analysis. Clearly the two courses of action are promoting each other.

By combining in analysis three some of the species as suggested by the first analysis, the overall picture is much improved. However, new species associations now become apparent. The canonical variate plot (Fig.8), is not directly comparable with the earlier plots, as the precision of the analysis in terms of distances is now less, with the D^2 values being considerably smaller overall as a result of an increase in the within groups dispersion. The jackknife misclassifications are now much reduced in number, with only three (2, 46, 52) involving OTUs that were not deemed outliers from the earlier studies . Two of the OTUs (6, 37) that were misclassified earlier are now translated into their 'correct' species. When OTU groups are combined in this manner one might expect some general improvement in the overall picture. However, one could also expect the appearance of some random associations between OTUs from the smaller groups and the centroids of the now larger more stable groups. This latter result does not in fact occur and thus further exemplifies the robustness of the *R.glabra*, *R.monroi*, *R.'sp.M'*, *R.haasti*, *R.tenuicaulis* and *R.subsericea* groups. General interspecific associations between the *R.parki–R.hookeri* complex and *R.haasti*, and between *R.hookeri an.–R.beauverdii* complex and *R.monroi*, *R.'sp.M'* are now evident on the canonical variate plot (Fig.8), but there is a lack of supporting evidence for this from the jackknife classifications, (Table 10).

The final analysis, incorporating both amendments suggested by the first analysis (i.e. combining *R.parkii & R.hookeri as.*, *R.australis & R.hookeri h.* and *R.hookeri an. & R.beauverdii* and omitting OTUs 1,6,30,37,26 and *R.'sp. K'*) provides a clear, largely unambiguous picture of the subgenus. There are strong indications that the three complexes *R.parkii–R.hookeri as.*, *R.australis–R.hookeri h.* and *R.hookeri an.–R.beauverdii* are more closely related to each other than any of the other species are to any other species. So despite the fact that these three complexes have sufficient individual unity to warrant 'species' status they may well be considered together a unit of a higher taxonomic rank on the basis of these results. The three remaining misclassifications (2, 38, 44) are all the product of both the association of the OTU with an OTU

now omitted and the position of the OTU relative to the altered centroid of its reformed group. The magnitude of the total dispersion as measured by the total of the eigen values for each analysis provides some further insight into the four analysis. A reduction from 1373 to 812 from the original groupings to these groupings with the deletions and a reduction from 1373 to 484 from the original groupings to the amalgamated groups, implies a loss of overall discrimination beyond a level that might be considered to be theoretically reasonable as result of two such moves. The implication is that the overall B.S.S/W.S.S ratio is being reduced thus creating a system of less clearly punctuated groups.

While the final groupings derived above clearly fit the data, they are not the only groupings that might do this, despite the harsh criterion of the minimum jackknife D^2 method as used here. Deleting OTUs from a study to achieve a validated set of groups such as these is clearly not a desirable taxonomic procedure, and may in fact negate the purpose of the exercise. However the removal of OTUs that are shown clearly to be outliers can by removing bias on the group centroids, improve the perceived structure of the derived groups. If an investigator has sufficient faith in his or her data set to consider taxonomic changes generally, then any numerically justifiable removals clearly have some biological significance. Although the repercussion of such removals in these present analyses was not as great as the amalgamation of the original hypothesised groups, they were such that a number of apparently unrelated associations were seen to change. This situation although an apparent numerical quirk, is of much practical significance if one considers the problem of accurately defining a group (mathematically a geometric centroid) and then attempting to see the relationships of different OTUs to this group. If the group is defined so that it incorporates members beyond a reasonable range, then the association of specific OTUs to the group will be unduly affected, as a result of their association with any abberant OTUs included in the group definition. Nevertheless this analysis has shown that the detection of outliers is likely to be of biological interest but unless these outliers are extreme their deletion may cause more problems than they solve.

The omission of OTUs purely on the basis of the size of their relative minimum jackknife D^2s is clearly a process that should in the ideal situation be made after all other suggested changes. However, in the present work the small sample sizes resulted in the impacts of the two types of changes being inextricably linked. Only by performing both changes independently and in

conjunction was it possible to justify clearly the removal of OTUs 1,30 and 26 and to reassess the status of OTUs 6 and 37. In general, the outliers detected in these analyses were those indicated by the multivariate normal assessment. The two exceptions were OTU 26, which showed no apparent abnormality in the earlier study, and OTU 34, which was considered an outlier in the multivariate normal study, but despite showing some tendency to lie on the outskirts of firstly *R.hookeri as.* and then the *R.hookeri as.-R.parkii* group, fell slightly short of deletion in this analysis on the basis of a somewhat arbitrary cutpoint on the minimum jackknife D^2. It is likely in certain numerical studies that outliers detected only in a multi-group analysis belong to a group within the general scope of the study but not to a group which has been defined.

The fact that the two-dimensional canonical plots showing about 70% of the appropriate variablility did not always highlight similarities seen in the jackknife D^2s, and vice versa, implies that these plots can give a misleading picture, even when they portray a significant amount of the variability. This conclusion is supported from all four analyses when a comparison is made between the jackknife D^2 and the canonical variate plots. The indication is that the canonical variate plots, by being constructed to highlight large inter-centroid distances, can often give a confused picture of the more subtle relationships. However, when the jackknife D^2s indicated close species affiliations these were supported in the canonical variate plots. This would seem to indicate that canonical variate plots should be seen as pictures of a certain resolution, this resolution being determined by the larger inter-centroid distances. If greater resolution is required to detect the exact nature of closer associations between specific groups, then only these groups should be included and the analysis re-run. The need to do this implies a lack of homogeneity in both within and between group dispersions, which itself implies a need for some taxonomic changes.

The lack of congruence between the jackknife classifications and the non-jackknife classifications is largely a product of group size, but it does suggest that many of the changes suggested here are 'fine tuning' and may not in fact be supported by characters of a different type.

An interesting side note to the above analyses concerns the performance of the ratio variables, which are often considered to add little new information to an analysis. In the data set used for these analyses the ratio variables comprise approximately one third of the total variable number and yet their proportion among the variables that failed the tolerance test at any one stage

ranged from 3/8 to 3/13. Thus they are apparently not over represented in this group, and appear to contribute additional useful information to the data set.

CONCLUSIONS:

The *a priori* classification for *Raoulia* subg. *Raoulia* (Ward, 1982) with the single OTU of *R.cinerea* removed, has been tested with discriminant analysis. The primary aim of this discriminant analysis has not been discrimination *per se*, but rather to generate an ordination output which is superior to a traditional hierarchical dendrogram, and which elucidates group composition and inter-group relationships given an existing classification. Results from a first such analysis suggested certain changes to group composition and the deletion of individual OTUs. The two types of alteration, suggested primarily by the relative magnitudes of the individual jackknife D^2 (OTU-centroid), were made individually and in combination, so that the individual impact of each was seen. The repercussions of OTU deletions were larger than expected, with some general changes occurring that involved more than the groups directly involved. The amalgamation of three pairs of groups produced a much more stable configuration that itself indicated different inter-group associations than were first revealed. Despite the harsh criterion of the minimum jackknife D^2 when sample sizes are small (n=3 to n=26) and variable, the limited number of misclassifications and their consistency generally implies that one is now dealing with the very fine detail of a classification that is in the main robust. A need for care is indicated when interpreting canonical variate plots, even when they contain up to 70% of the variability, without cross validation to the jackknife D^2s. The results also imply that an overemphasis on non-jackknife D^2 when sample sizes are small is likely to lead to conservative and possibly incorrect interpretations of the data, as is an analysis that focusses solely on misclassifications, jackknife or not, rather than on the magnitude of the individual D^2s (OTU-centroid).

CHAPTER 5

COMPARATIVE ASSESSMENT OF GROUPING STRATEGIES USING ALL CHARACTERS

"The poet's eye, in a fenzy rolling,
Doth glance from heaven to earth, from earth to heaven;
And, as imagination bodies forth
The forms of things unknown, the poet's pen
Turns them to shapes, and gives to airy nothing
A local habitation and a name..."
WILLIAM SHAKESPEARE

INTRODUCTION:

The results from a discriminant-classification type analysis as described in the previous chapter have two major weaknesses if one's aim is to derive the correct underlying OTU structure rather than to test a classification.

1)The results are very much biased in the direction of reproducing the groups as defined *a priori*. This is not to say that significant changes cannot be suggested and compared by a discriminant analysis, but, regardless of whether the hypothesised groups are vindicated or not by the data, the discriminant analysis will not necessarily lead to an optimum grouping strategy. Minimising misclassifications, for example, will lead to a solution, but this would not necessarily be an optimum one.

2)The standard discriminant analysis cannot incorporate multi-state nominal characters unless they are recoded as a set of highly correlated binary characters. Further to this, if any characters are invariant within all groups, i.e., they are useful diagnostic characters, they cannot be included in the analysis. Thus many characters of taxonomic importance are not being utilised in the standard discriminant analysis. Their influence on grouping stategies is thus lost.

In overcoming the above two problems a number of different methods have been proposed as 'stopping rules' for the optimisation of the number of groups derived from an exploratory analysis. The methods rely mainly on the output from a sequential or non-sequential cluster

analysis as the guideline for group construction, and in so doing they obviate the need to calculate and assess all possible grouping combinations. Annotated descriptions of many of the 'stopping rules' can be found in Everitt (1974, 1979) and Hill (1980b).

One such measure, first derived by Ratkowsky & Lance (1978), is an extension of a measure applicable to nominal characters (Cramer, 1946), so that characters in a numeric form may also be used. This measure assumes a set of independent Euclidean dimensions so that it may be calculated for each character independently. These values are then totalled and averaged to give an overall measure for the current grouping strategy. The calculation for nominal characters is :

$$C_i = ((\chi^2/n)/(\min.(m_i-1,g-1)))^{1/2}$$

where: χ^2 = standard chi-square measure of independence

n= total number of OTUs

m_i= number of states for character i

g= number of groups

and for numeric characters:

$$C_i = (BSS/TSS)^{1/2}$$

where: BSS=between groups sums of squares

TSS=total sums of squares

In calculating C in this manner, it is constrained to lie between 0 and 1. Allowance is made for changing group number as different grouping strategies are assessed and compared, by dividing the average measure overrall characters by $g^{1/2}$ to give the final measure of the congruence of the data with the hypothesised groupings. A maximum value for the measure is taken as being indicative of an optimum grouping strategy. The measure was successfully applied to a number of different data sets (Ratkowsky & Lance, 1978), but has subsequently been discredited (Hill, 1980b; Ratkowsky, 1984). It was discredited on two important grounds:

1)In order to determine the groups for calculating C, a phenon line was seen as necessary. The phenon line has never been recorded as being accepted as an accurate means of determining group composition, and Hill (loc. cit.) outlines the obvious problems involved in attempting to use it for this purpose.

2)The measure incorporating a divisor of $g^{1/2}$ rather than say, $(g-1)^{1/2}$, was derived not objectively but empirically. In a paper by Ratkowsky (1980), the author admitted that the measure tended to produce a small number of groups, and this was shown by Hill (loc. cit.) to be a direct result of the decreasing maximum of the measure ($1/g^{1/2}$), as a product of increasing group number. Thus the use of the magnitude of C as indicative of grouping success without reference to the changing maximum, dependent on group number, could not be maintained.

These two criticisms of the measure are valid as they stand and both Hill (loc. cit.) and Ratkowsky (1984) have modified the measure to arrive at new methods for determining the optimum grouping strategy. The complete departure from the original form by Ratkowsky (1984) and the continuing problem that he has experienced in attempting to achieve an objective derivation of an appropriate divisor, make his amended 'stopping rule' unattractive to anyone who may have been attracted by the uncomplicated intuitiveness of the original form. Hill's (loc. cit.) revision answers both criticisms of the earlier measure but relies on the binary fusion of groups as represented in the standard dendrogram.

If any form of OTU reallocation is used in an attempt to improve the groupings, and the number of groups is not altered in the process, the original form of the measure with any power function of g as the divisor will suffice. In this instance the advantage of having a direct access to the individual character contributions remains. However, if one wishes to retain the general form of the measure and view the possible fusions and splittings of the groups in a less strict, non-hierarchical sense, there is no available measure. Major modifications to the original method are thus proposed. These modifications allow the probabilistic assessment of changes, in group number and in variable number, on an overall measure of the grouping success.

METHODS:

One problem, in developing a technique that accumulates individual variable results for different variable types into an overall measure, is to ensure that each variable is standardised to contribute with equal weighting. The Cramer measure (Cramer, 1946) does this for nominal characters of a varying number of states by incorporating:

(i)the minimum degrees of freedom (d.f.) of either the number of states or the number of groups

(ii)the total number of OTUs and

(iii)the χ^2 value

into the calculation, so constraining the result to lie between 0 and 1. For nominal variables alone, an alternative to this procedure is to calculate the Chi-square statistic for each variable independently, and then, using the independent status of each χ^2, sum these and their respective d.f. to give an overall χ^2 value with d.f. The total d.f. will always equal $(g-1)*(\Sigma m_i-p)$. In this procedure the size of any individual χ^2 contribution is standardised by its associated d.f. Given this calculation, any change to an overall χ^2 value brought about by a change in group number or a change in character number can be compared to the χ^2 distribution, with the difference in d.f. for the two measures being the relevant d.f.. Any absolute measure other than a difference is likely to be very significant statistically, in the same manner as a Wilks lambda may frequently be when one is performing a MANOVA. This value will not be very rewarding on its own, and the null hypothesis associated with it will be of little interest. However, the null hypothesis of a non-significant gain or loss to a measure of grouping 'success' is of much interest. This technique of assessing the difference between two related values for a given test statistic is analogous to one described by Rao (1947, 1950) where the change to a specific D^2 value based on p characters and one based on p-q characters was interpreted probabilistically.

It is apparent that the value for the d.f. cumulated in the above manner for any example in numerical taxonomy, will make a direct comparison with the χ^2 distribution via conventional χ^2 tables impossible. In these situations (d.f. >100) one can use the normal approximation to the χ^2 distribution i.e.

$$Z=(2\chi^2)^{1/2} - (2v-1)^{1/2}$$

or for d.f. >30

$$Z=((\chi^2/v)^{1/3} -(1-(2/9v)))/(2/9v)^{1/2}$$

where v=d.f..

When the value of the normal deviate is negative and significant and one has combined groups or deleted characters, then clearly such a change has resulted in a 'better' grouping as determined independently by all the variables. Alternatively put, the significance of the χ^2 statistic for the combined groups is greater than for the original groups. A positive result for the normal

deviate in identical circumstances indicates that combining the groups is not supported by the data. The null hypothesis being addressed here is that manipulations of the groups are neither supported nor contradicted by the data. A rejection of this null hypothesis on the basis of a significant χ^2 has then to be interpreted as support for or opposition to the revision. If the groups are split or more variables are added so that there is an increase in the d.f., the same null hypothesis applies, but the interpretation of a rejected null hypothesis is reversed. A negative normal deviate indicates a lack of support for the change, while a positive deviate indicates that the change is supported by the data. If the null hypothesis is not rejected so that the results are inconclusive either way, other 'costs' or 'benefits' involved in certain grouping strategies have to be considered before a decision can be made.

It seems possible that there may in certain circumstances be a change in the χ^2 value that is opposite in direction to the change in the d.f.. In this situation the result should be considered as being indeterminably significant beyond the level for no change in the χ^2.

The problem of how to incorporate an appropriate and compatible measure for the numeric characters still remains. One solution to this is to calculate the F-ratio for these characters, as for a one-way ANOVA, and given that the error d.f. (n-g-2) is greater than 60, this ratio when multiplied by the group d.f. (g-1), can be compared with the χ^2 distribution with g-1 d.f. In this manner each numeric character can produce a χ^2 value with g-1 d.f., both of which may in turn be added with or without the contributions from the nominal characters to yield a single χ^2 value and d.f.. This approximation of $F_{(g-1),(n-g-2)}$ by a $F_{(g-1),\infty}$ when n-g-2 is not >120 is likely to lead to an increase in the probability of a type I error. The percentage discrepancy between $F_{m,60}$ and $F_{m,\infty}$ at $\alpha=0.05$, varies from 4.5% greater at m=1 to 39% greater at m=∞. As g-1 is not likely to exceed 20 in most numerical taxonomic examples, the maximum discrepancy at $\alpha=0.05$, becomes 11.5% per numeric character. There are clear indications that absolute values and differences that lie close to the rejection region would have to be considered conservatively. With fewer groups and more characters the approximation will improve, but as α increases the precision of the approximation decreases. In the examples that follow, with g-1 approximately 12 and n-g-2 approximately 72, the level of α for $F_{12,\infty}$, using $F_{12,72}$ ($\alpha=0.05$), is 0.0307, for $\alpha=0.01$ it is 0.003. As one is most likely to be looking at the differences in χ^2 values rather than absolute values, and if the error in the approximation is approximately equal in the two χ^2 values, then these errors

may be cancelled out or at least reduced in the subtraction. In an attempt to test the accuracy of the approximation and to give 'exact' probabilities for the change in χ^2 values a PASCAL computer program (Appendix D) has been written that calculates F and χ^2 tail probabilities and χ^2 values given tail probabilities. This program incorporates the algorithms of Ibbetson (1963), Morris (1969) and Hill & Pike (1967).

The removal of individual OTUs or OTU sets that do not alter the number of groups has no effect on the d.f. and thus cannot be probabilistically assessed using the above methodology. Nevertheless, the differences between the normal deviates calculated with and without any particular OTU or OTU sets should always be negative, if the OTU(s) disturbs rather than contributes to the defined grouping structure. The one-sided null hypothesis being addressed is that the OTU belongs within its hypothesised group, so that unless a large negative difference between the normal deviates is shown, there is no evidence for excluding the OTU(s).

The above methodology (C.A.G.S.: Comparative Assessment of Grouping Strategies) will be used to test changes to the classification of the subgenus as suggested by the previous analyses. Thus the individual OTUs highlighted as being outliers from the subgenus on the basis of the numeric and binary characters, will be excluded and the individual impacts on the overall χ^2 values assessed. The amalgamation of groups, as suggested by the discriminant analysis, will be made and the changes in the χ^2 values will be assessed probabilistically. The approximation of $F_{g-1,g-n-2}$ by $F_{g-1,\infty}$ will be verified for one grouping stategy against the baseline grouping. To assess this approximation accurately it is necessary to extract the variables whose F values from both grouping strategies do not return α levels that are smaller than the likely accuracy of computation, i.e. 10^{-6}. To test the approximation, the difference between the χ^2 values, as calculated for the two grouping strategies, will be computed in two ways:

(i) Using the approximation as above and

(ii) using the α levels from the $F_{g-1,n-g-2}$ to calculate 'exact' F values for (g-1) and (∞) degrees of freedom, and then summing these F values to give a further two χ^2 values from which a second difference can be extracted. A comparison of these two values will not only give some measure of the bias in any single χ^2 value, but will also give some indication as to the degree to which the bias is being cancelled out in the subtraction. Clearly this method of determining the

bias is itself dependent on the similarity of the two F distributions and is thus subject to some degree of error. Nevertheless it is the most attractive method of removing a significant proportion of the bias to assess the impact of this same bias.

For all these analyses *R.cinerea* will be excluded, thus n=85 and g=13 for the 'baseline' grouping strategy.

RESULTS:

The z-score for the overall baseline measure for the hypothesised 13 groups and 85 OTUs was 100.97, calculated from a χ^2 of 23233232 with 1656 d.f.. The z-scores for the respective and independent deletions of OTUs 1,6,26,30,34 and 37 were 100.21,101.43,100.96,100.58,99.8 and 100.33. Only one of these differences, -1.46 for OTU 6, gives a negative difference from the baseline and this is not of sufficient magnitude to suggest that the OTU is causing any significant disruption of the groups as defined in the 'baseline'.

Table 12 lists the different combinations of the original groups and the loss or gain of such strategies in terms of the χ^2 measure. Only the amalgamation of *R.beauverdii* and *R.hookeri an.* is supported by the χ^2 of the difference as shown in Table 12. If this amalgamation were to proceed, none of the further amalgamations or deletions considered are supported.

Using the lowest 15 F values from the analysis of the baseline grouping, a cumulative χ^2 of 390.72 with 180 d.f. was calculated using the approximation of $F_{g-1,n-g-2}$ with $F_{g-1,\infty}$. When this value was recomputed using the 'exact' probabilities to compute the precise $F_{g-1,\infty}$ values, the χ^2 was reduced to 351.58. The same procedure using grouping strategy 4 gave χ^2 values of 374.88 and 336.7 respectively, both with 165 d.f.. The two differences, either of which could be used to determine which strategy is superior, were 15.84 and 14.88 respectively, both with 15 d.f..

DISCUSSION:

It has been the dual purpose of this set of analyses to test an intuitive and yet objective measure of how well a data set of mixed variable types supports one given classification over another, while simultaneously testing the appropriateness of changes to the classification that were suggested by the previous analyses involving only numeric characters.

TABLE 12: THE CHANGES IN X^2 VALUES AND Z-SCORES ASSOCIATED WITH DIFFERENT GROUPING STRATEGIES.

STRATEGIES	$X^2_1 - X^2_2$	D.F.$_1$ - D.F.$_2$	Z
1 Vs 2	365.7	138	10.5
1 Vs 3	2980.7	138	60.6
1 Vs 4	14.9	138	-11.1
1 Vs 5	144.4	138	0.4
1 Vs 2,3,4	3309.3	414	52.6
4 Vs 2,4	361.2	138	10.3
4 Vs 3,4	2991.2	138	60.8
4 Vs 4,5	146.8	138	0.5

STRATEGIES

1) The original groupings
2) Combining *R.parkii* and *R.hookeri as.*
3) Combining *R.australis* and *R.hookeri h.*
4) Combining *R.hookeri an.* and *R.beauverdii*
5) Removing *R.'sp.K'*

Without specific reference to the taxonomic repercussions of these analyses, it is clear that the measure itself is giving results that are consistent and potentially useful. The amalgamation of *R. beauverdii* and *R.hookeri an.* was consistently favoured as a strategy, while the current status of all other species was equally consistently supported. In a similar fashion the results concerning the omission or not of *R.'sp.K'* were indecisive. It is apparent from these results in conjunction with the previous chapter that the individual identity of many of the species, notably *R.parkii*, *R.australis*, *R.hookeri h.* and *R.hookeri as.* is greater if nominal characters as well as numeric ones are considered.

Although this analysis does not specifically deal with the character contributions to the individual χ^2 values, it is of interest to note that in all analyses greater than 50% of the characters with the 10 largest contributions to each of the χ^2 values were nominal characters.

The greater integrity of the original groups revealed in the comparisons between different grouping strategies is mirrored in the lack of strong evidence for the deletion of any the previously considered outliers. Only the deletion of OTU 6 gave an improved grouping structure, but the magnitude of this improvement precluded its deletion.

The problem of the approximation of an $F_{g-1,n-g-2}$ distribution with an $F_{g-1,\infty}$ is obviously of some concern. The p-values for the two χ^2s calculated from the lowest 15 F values are clearly too small to compare accurately. However, the differences between the χ^2 calculated in both manners are of interest. The difference between the individual Chi-square values for both grouping strategies is approximately 10% and for the difference between the χ^2s this falls to about 6%. There is thus some indication that the bias is reduced in the final measure on which an optimum grouping strategy is determined. In this instance there was no evidence of the superiority of one strategy over another in either of the differences $p=0.393$ and 0.460 respectively. However, when the differences are larger the p-values are thus greater, and the precision of this approximation will decrease. It is clearly invalid to approximate one distribution to another when the distributions are of a very different form. In the above situation, $g=12$ and $g=11$, the $F_{g-1,n-g-2}$ distribution is not greatly different from $F_{g-1,\infty}$. Thus the approximation may be used provided that one is conservative when considering p-values that are near to the rejection region. The approximation will obviously improve as the number of groups decreases and the number of OTUs

increases. Thus, when strategies are compared, there is likely to be some bias towards not supporting the strategy with fewer groups or more OTUs. The bias in the approximation is clearly to increase the probability of a type I error, and the more strongly any strategy with more groups or less OTUs is favoured over another, the greater this bias will be. There is likely to be a dilution of this error by the inclusion of a large proportion of non-error-prone nominal characters in the calculation of the X^2 values. This specific effect has not been addressed. The likely error introduced by the approximation would not have been large enough to reverse any of the above decisions concerning the grouping strategies. Clearly in other situations this may not be the case, and a conservative approach would need to be taken.

From a taxonomic perspective the results obtained from these analyses are rewarding. The 'outliers' suggested by the methods involving only numeric characters were not considered to be unusual beyond their species' ranges by a specialised taxonomist, (Dr J.M. Ward, pers.com., 1987). There was supporting evidence which implied that these OTUs were peculiar in respect of their states for the numeric characters, e.g., OTU 34 appears from the herbarium sheet to have a different overall form which may be attributed to its youth, while OTU 30 is a coastal form of *R.hookeri h.* and has a larger, more gross appearance than the other OTUs of this species. Nevertheless, it would appear from this analysis that these unusual features do not extend to their diagnostic nominal characters.

The hypothesised amalgamation of species as suggested by the discriminant analyses, were supported to a degree by taxonomic evidence in that the close, inter-species relationships shown are believed to exist. However, apart from the uniting of *R.beauverdii* and *R.hookeri an.* the other possible combinations were not considered appropriate at the species level. The combining of *R.beauverdii* and *R. hookeri an.* however, is one that is supported by their treatment in Allan (1961, p.705)

CONCLUSIONS:

A new measure has been proposed to test the degree to which any variable set supports one grouping strategy of OTUs over another. This measure relies on the summation of the independent

character contributions for each character. The measure requires an approximation of a $F_{g-1,n-g-2}$ value with an F with g-1,∞ d.f., and this approximation is likely to lead to a degree of bias that increases the probability of a type I error in the independent character contributions. It thus leads to a level of unreliability in the final difference measure. On the basis of this a conservative approach to the final interpretation is advocated. The conditions for an improved approximation include fewer groups and more OTUs. The measure has been shown to give biologically useful results on the particular data set used.

CHAPTER 6

DIAGNOSTIC CHARACTERS

"What's in a name? that which we call a rose
By any other name would smell as sweet."
WILLIAM SHAKESPEARE

INTRODUCTION:

This chapter relates to what is usually the final stage in the taxonomic sequence, which is to generate a key for the ready identification of the groups concerned by the extraction of diagnostic characters. Many different algorithms and programs exist for this purpose (e.g., Pankhurst, 1971; Dallwitz, 1974; Payne, 1975). Such programs generally rely on a pragmatic, error-free approach that successfully classifies the given data set, with generalisations appropriate only in relation to the sampling technique used to generate the data. This means that clearly defined taxa are frequently used rather than those determined by the preceding techniques, which involve a hypothetical rather than a definitive classification. They are, however, extremely flexible in that indices of character reliability and character cost can be user supplied, they minimise the length of the key, and missing data can be allowed for. These programs have not as yet entered into the field of continuous numeric characters *per se*, so that techniques for optimising cutpoints in these characters for diagnostic purposes remain unexplored. The algorithms tend to use binary characters preferentially and then multi-state characters as required.

The techniques used in the accepted programs (Dallwitz, 1974; Payne, 1975) provide useful working keys and avoid the probabilistic models of such techniques as discriminant analysis and multinomial analysis which are frequently inappropriate in a taxonomic sense. In certain circumstances there may be a need to use elaborate statistical modelling to accurately discriminate groups, but unless such models have an empirical justification their use beyond the sample data set is likely to be limited and their practical application cumbersome.

It may well be that keys generated with no recourse to an optimum key and in the manner of repeated sequential binary subdivisions are not the most efficient when the final sequence is determined. However, if satisfactory working keys are generated in this manner there is no real need for more elaborate techniques that incorporate some form of recursive optimisation.

METHODS:

Given the 'final' thirteen species groupings, and indications of the independent discriminating ability of the discrete characters from the output of the previous chapter, key characters will be identified and used to form an identification key. This key will be constructed in a dichotomous manner, with the hierarchy of the divisions determined by the maximum-minimum multivariate F-ratios between species. These ratios will be extracted from a discriminant analysis output generated using the 12 species with more than one OTU. In this manner the successively less distinct species will be isolated and split from the remainder of the data set. If a complete key can be constructed in this manner with discrete characters, the resultant vector of scores for each group is likely to be the same as that generated by a multinomial model, given that there is only one combination of characters that will differentiate all groups. However, given the comparatively small number of groups, this result will be achieved with considerably less computer time and more direct contact with the data set than with conventional computer programs. This direct contact makes it possible to select character combinations that incorporate characters that are readily available and have easily discernible states: in effect one is applying an intuitive weighting to the character usefulness, a technique that Dallwitz (1974) applies numerically. If no single discrete characters or discrete character combinations can be readily isolated for discrimination, the possibility of using continuous characters with appropriate cutpoints will be explored. Failing this, discriminant functions will be generated for specific dichotomies. If these provide accurate discrimination and an empirical interpretation of the function(s) can be made, they will be used for diagnosis.

The characters that determined the outlying status of several OTUs in the studies involving numeric and binary characters will also be identified. Given that this status is relative to the subgenus and not to the individual species, the most efficient method of isolating these characters is to calculate the z-scores, relative to the means of the subgenus, for each of the OTUs for each

binary or numeric character. Given that n=86, a z-score exceeding 2.0 is considered significant in this context.

RESULTS:

Table 13 lists the characters used for the sequence of subdivisions for each species. Apart from separating *R.australis* from *R.hookeri h.*, each of the dichotomies could be made using a unique character state for a minimum of one character. The final distinction of *R.australis* from *R.hookeri h.* could be made only by using the combined states for two variables. If the three OTUs within *R.australis* known to be hybrids are removed, the distinction can be made on the number of leaf traces.

Fig.10 shows the canonical variate plot for the 'final' grouping strategy. The F values associated with the inter-centroid D^2s were used to determine the sequence of subdivisions for the key.

The characters that had outlying values for the OTUs identified as 'outliers' are listed in Table 14. OTU 26, which the discriminant analysis showed as being 'unusual', must have a peculiar set of character combinations rather than outlying values for any character. A similar but less extreme situation exists for OTUs 6 and 37. There are indications that an intuitive correlation exists between the magnitudes of the z-scores and the number of characters with large z-scores, and the size of the D^2 case to centroid. From the 63 numeric characters used here the most extreme values for 22 are to be found in four OTUs.

Interpreting the character combinations listed in Table 14 in biological terms is not straightforward. 'Spurious' measurements are inevitably going to cloud such attempts. However, some generalisations can be made. The characters with extreme values for OTU 1 are largely related to a large pappus and unusual corolla dimensions. OTU 30 appears to have a large leaf base and leaf lamina and, again, unusual corolla dimensions. OTU 34 has a broad lamina, large achenes and wide corolla breadth. Apart from the diagnostic characters OTU 77 has large leaves, a generally unusual leaf form, a large number of inner phyllaries and a high minimum number of florets.

TABLE 13: KEY CHARACTERS FOR THE FINAL SPECIES GROUPS

SPECIES	CHARACTERS	STATES
R.cinerea	1) pappus hair number*	<33
	2) pappus series*	single
	3) pappus hairs shed*	singly
	4) pappus hair width*	>0.4mm
R.tenuicaulis	1) long narrow papillae borne at a wide angle	yes
R.haastii	1) vein order	3
R.glabra	1) leaf mucro form*	upturned
	2) tomentum coverage lower side*	glabrous except at centre
	3) texture of basal part of pappus hair	distinct from upper part
	4) lamina tomemtum density lower side*	absent
	5) receptacle curvature*	conical
R.'sp.M'	1) expansion of pappus hair tip*	great
	2) shape of leaf apex	subacute
	3) lamina folding	none
	4) shape of phyllary apex	acute
	5) leaf mucro form	absent
R.monroi	1) shape of main vein form	curved
R.hookeri an.	1) inner phyllary colour	dark brown
R.'sp.K'	1) main vein as percent of lamina breadth	greater than 4%
	2) expansion of pappus hair tip	none
	3) pappus hair tapering to tip	yes
R.subsericea	1) highest vein order	2
	2) number of leaf traces	3 strong
R.parkii	1) lamina folding	boated at tip only
R.hookeri as.	1) inner phyllary colour	fulvous to gold to tan or mid-brown
*R.australis***	1) apex of pappus apical cell and shape of phyllary apex	sub-acute rounded
*R.hookeri h.***	----	----

*indicates a state unique to this species within the subgenus

**For an explanation see first paragraph of the results section (page 79)

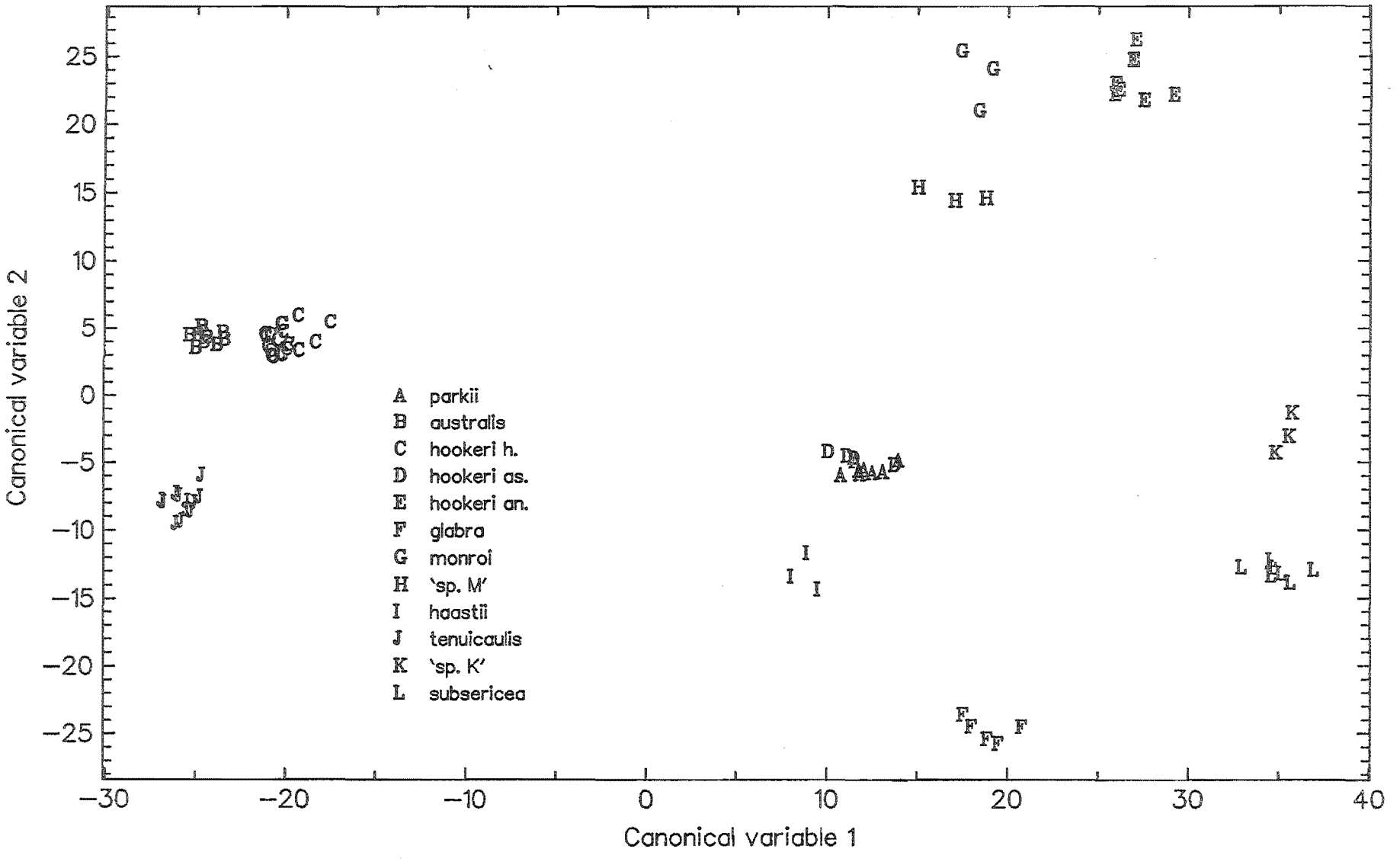

Figure 10: Plot of canonical variable 1 against canonical variable 2 for 12 species groups.

TABLE 14: ABNORMAL CHARACTERS FOR OUTLYING OTUs

OTU	CHARACTERS WITH Z >2.00
1	55*,76*,77*,79,80*
6	56
26	nil
30	8*,10,11*,12,50*,78,79*,89*
34	8,43*,44*,45*,46*,49*,50*,85*,86*
37	53*,72,92
77	7*,9*,10,13,23,63,65,90*,91*, 92*,94*

*indicates an extreme value for the subgenus

DISCUSSION:

The results listed in Table 13 provide a clear indication that the large number of characters and the small number of groups in this data set have combined to make the construction of an identification key a relatively straightforward procedure. No more elaborate technique than that used here was required. The method of identifying the most different species via the maximum-minimum F-ratios may clearly have helped, but it must be remembered that many of the diagnostic characters selected were multi-state discrete characters that were not part of the calculation of the F-ratios. The early detection of 'useful' characters (that is characters with a genuine practical diagnostic use) as a by-product of the work described in the previous chapter may also have saved a considerable amount of time. The procedure of selectively splitting one species from the others will ultimately give a unique diagnostic vector for each species, but obviates the need to specifically seek this, as would be necessary if the method involved the instantaneous separation of each species from all others. The final diagnostic tree here developed is not an optimum one. Clearly where unique states for two species were not identified until after other divisions had been made, the technique is not perfect. Nevertheless, these unique states have been retrospectively identified. The key could now be rearranged into an alternate and more efficient one.

The identification of outlying character states for the outlying OTUs provides useful information on the nature of these OTUs (Table 14). If one is considering only numeric characters then the existence of extremes for individual OTUs on individual characters could well be the product of measurement error, or some phenotypic phenomenon. On the other hand the existence of unusual character combinations, potentially revealed in a discriminant-type analysis or in an analysis of the D^2 OTU to centroid, and with little evidence of unusual individual states, is more likely to be of some taxonomic significance. If both forms of this exist, as in the case of OTU 77, then the OTU is distinct. It is possible that via either the D^2 OTU to centroid or the z-scores for individual character states one is detecting OTUs that are of an extreme magnitude for one or more characters, e.g., OTUs 30 and 34. Clearly unusual character combinations may be present that imply an unusual position for an OTU within the multivariate hyper-sphere delimited by the taxon being studied. However, such 'outliers' are only likely to be detected via some technique of dimension reduction, such as discriminant analysis. Such 'outliers' are not to be considered removed

from the taxon being investigated but are clearly of much interest in relation to the intra-taxon variability. OTU 26 is one clear example of this.

CONCLUSION:

A clear, unambiguous identification key has been established for the 'final' 13 species. The construction of this key, by hierarchical binary divisions, has been shown to be appropriate in this situation, but this method will not necessarily generate an optimum key. The small number of groups and the large number of 'useful' discrete characters made it unnecessary to use cutpoints on numeric characters. A knowledge of the characters' grouping capacity and the relative distances among the group centroids is indicated as providing useful information for the easy construction of a key, but this has not been specifically tested.

The understanding and interpretation of unusual OTUs has been shown to be improved by the analysis of individual variable distributions. OTUs that are thought to be unusual on the basis of earlier studies may well be explained in terms of non-taxonomic variability as revealed by the magnitudes of independent character z-scores. When unusual OTUs do not have significant z-scores, their status is likely to be the product of specific character combinations, implying either a new group within the taxon being studied, or a group external to the study. Discriminant analysis is likely to detect both forms, whereas the individual D^2 OTU to centroid will detect only the latter form.

CHAPTER 7

CONCLUSIONS

"Life is the art of drawing sufficient conclusions from insufficient premises." SAMUEL BUTLER

The purpose of this thesis has been to explore specific avenues within each phase of a systematic study involving numerical taxonomy. It is anticipated that the results from this thesis, rather than being conclusive in themselves, will further add to the pool of worked examples that eventually vindicate or repudiate specific methods. Clearly the applications of all the methodology portrayed here lie broadly in the field of multivariate description and not to numerical taxonomy alone.

The first section dealing with character selection has shown that the inclusion of additional derived characters *per se*, in this instance ratios, is a justifiable action in terms of adding sufficient useful information to a data set. Ratios are still shown to have irregular distributional patterns although a technique for minimising this is advanced. However, given that univariate distributions are infrequently considered in multivariate studies and that irregular distributional patterns were subsequently shown to have little influence on the multivariate distribution, these patterns are not thought to be harmful, in contrast to the conclusions of Atchley et al. (1976).

The second section deals with the statistical pre-requisite for many multivariate analyses, multivariate normality. The findings of Reyment (1971) are generally supported and are extended. The importance of this assumption is shown to have biological as well as statistical repercussions and thus the testing of this assumption is advocated as a preliminary step in multivariate analyses. Outlying OTUs rather than irregularly distributed characters are shown to be the primary influence on multivariate normality in this instance, although the two interrelated effects of the sampling scheme, character number and OTU number, are shown to have a significant bearing on the statistical formation of a uniform multivariate distribution. In creating a taxon with a multivariate

normal distribution over the sampled characters by the deletion of outlying OTUs, it is shown that one is statistically defining the taxon being studied and thus comparisons with other taxa and within the taxon can accurately be addressed.

The third section, although utilising an *a priori* defined set of groups, is essentially an exploratory study using discriminant an alysis. In the course of this analysis, various alterations to the grouping structure were suggested and then implemented. The success of these moves was judged on the basis of the magnitude of the minimum jackknife D^2 of each OTU to any particular species' centroid. This technique is seen to be sensitive to sample size but is less conservative than the assessment of non-jackknife D^2. The highlighting of OTUs on the periphery of the subgenus in the previous chapter enabled these to be followed through a typical discriminant analysis. The results indicate that large minimum jackknife D^2 should be explored as being as important as the frequently highlighted misclassifications. However, some of the outliers were still shown to provide robustness to their hypothesised groups.

The fourth section attempts to redress the balance away from numerical decisions based purely on binomial and continuous characters, by developing a technique (C.A.G.S.) whereby suggested changes within the taxon could be assessed probabilistically using the entire character set. This technique is established as being more intuitive and yet less subjective than the techniques of Hill (1980b) and Ratkowsky (1984).The technique is likely to be inappropriate when the sample size is small. Nevertheless in this particular example with an 'intermediate' sample size (85), the results were biologically meaningful and useful. The method is therefore advocated for further experimentation.

The final chapter in the sequence, using previously gained information on the inter-centroid distances and the independent discriminating ability of each character, generates a working identification key for the 'final' taxonomic groups. The technique used was logical and uncomplicated and while it produced the desired effect without any recursive methodology, it is not likely to generate an 'optimum' key in any circumstances.

At each stage in the above procedure the purpose has been to improve the understanding of the interelatedness of parametric statistics and taxonomy by investigating the biological significance of appropriate statistical results. Frequently the interface between the theoretical dimensions of

statistics and the applied work of biological researchers is neglected. As Gower (1988) expressed it, "...what we must be careful not to do is to frighten away those more interested in applications than theory, by putting too much stress on mathematical and statistical topics. Provided the Federation [The International Federation of Classification Societies] and its members retain a balance between these two aspects of its work I believe it will have a long and profitable future that will benefit both practitioners and theorists". The work of mathematicians and physicists of previous generations has shown that all inanimate objects are restricted and patterned in their deviations from the general tendency towards entropy by the fundamental laws of physics. The problem confronting biologists today is the mathematical descriptions of both the laws of nature that determine the development and evolution of living organisms and the mathematical definitions of the organisms and groups of organisms that develop and evolve. The mathematical definition of individuals, populations, species etc. and the relationships with other individuals, populations and species is requisite if a non-teleological understanding of living organisms is to be gained. Clearly the accurate mathematical definition of an organism is of primary importance if comparative statistical methods are to be used and it is the development of this definition that currently presents the greatest challenge. Frequently statistical methodology is held at fault for 'inappropriate' results, when the methodology may have been vindicated if 'appropriate' results had been generated. This is not a problem involving the methodology or the definition of 'appropriate' but rather one that hinges on the accurate representation of the field of study in a numeric form. Once this not insignificant problem is overcome the levels of variability for individual characters will unambiguously define phenotypic and genotypic characters. As a consequence of this the core of correlated characters that uniquely defines and restricts an 'evolutionary unit' will be able to be determined. The punctuated point at which this 'unit' ceases to be a single evolving entity corresponds to a change in level within the taxonomic hierarchy and the individual entities (units) that have evolved from this are obviously at a lower level in the hierarchy. These derived groups will contain the core of correlated characters unique to their progenitors but will have additional 'evolved' characters that distinguish them from other groups with the same progenitor. These 'evolved' characters are those eagerly sought within the field of cladistics. It is in the identification of the core character sets and thus the multivariate level of variability between groups of similar and different levels in the hierarchy that the strangely divergent fields of phenetics and phylogenetics will inevitably unite. The capacity to discern these character sets is available today

without the need for a genetic definition of individuals. The mathematical knowledge is also currently available; all that is required is considerable manpower and improved inter-disciplinary discussion. It is only hoped that in some small way this thesis has advanced the achievement of the genuinely 'natural classification'.

"Yet nature is made better by no mean
But nature makes that mean. So'er that art
Which you say adds to nature, is an art
That nature makes. You see, sweet maid, we marry
A gentler scion to the wildest stock
And make conceive a bark of baser kind
By bud of noble race. This is an art
Which does mend nature-change it rather; but
The art itself is nature."
WILLIAM SHAKESPEARE

ACKNOWLEDGEMENTS

I gratefully acknowledge the supervision of Dr J.M. Ward and Mrs H. Langer without whom this thesis would not have been possible. I also record my indebtedness to Dr J.M. Ward for the use of her data set, from which all the data for this thesis were drawn. I further acknowledge the extra-curricula support and advice from Dr E. Wells, Mr G. Findlay and Dr R. Sedcole. Thanks are also due to Dr D.G. Lloyd and Dr D. Kelly for helpful advice on the writing of this manuscript.

Much appreciation is also given to the following friends and colleagues who on many occasions furnished stout and essential support well beyond the bare call of duty; (apologies are given for any omissions) Bert, Frank, Dave, Dave, Jo, Leigh, Dick, John, Paul, Phil, Ian, and Matt.

The continued support and assistance (non-financial) from my family especially Katie, James and Andre is also gratefully acknowledged.

REFERENCES

ALBRECHT, G.H. (1978) Some comments on the use of ratios. Syst. Zool., 27: 61-68.

ALBRECHT, G.H. (1979) The study of biological versus statistical variation in multivariate morphometrics: the descriptive use of multiple regression analysis. Syst. Zool., 28: 338-344.

ALLAN, H.H. (1961) Flora of New Zealand. Vol. 1. Govt. printer, Wellington. 1085p.

ANDERBERG M.R. (1973) Cluster analysis for applications. Academic Press, New York. 359p.

ANDERSON, G.J.; WALBERG, H.J. AND WELCH, W.W. (1969) Curriculum effects on the social climate of learning: a new representation of discriminant functions. American Educational Research Journal, 6: 315-328

ATCHLEY, W.R. (1974) Morphometric differentiation in chromosomally characterised parapatric races of morabine grasshoppers (Orthoptera: *Eumastacidae*). Aust. J. Zool., 22: 25-37.

ATCHLEY, W.R. (1978) Ratios, regression intercepts and the scaling of data. Syst. Zool., 27: 78-82.

ATCHLEY, W.R. AND ANDERSON, D. (1978) Ratios and the statistical analysis of biological data. Syst. Zool., 27: 71-77.

ATCHLEY, W.R.; GASKINS, C.T. AND ANDERSON, D. (1976) Statistical properties of ratios I. empirical results. Syst. Zool., 25: 137-148.

BAGGALEY, A.R. AND CAMPBELL, J.P. (1967) Multiple discriminant analysis of academic curricula by interest and aptitude variables. J. Educ. Meas., 4: 143-149.

BARTLETT, M.S. (1947) Multivariate analysis. J. Roy. Stat. Soc. Supp., 9: 176-197.

BIRKS, H.J.B. AND PEGLAR, S.M. (1980) Identification of Picea pollen of late quaternary age in eastern North America: a numerical approach. Can. J. Bot., 58: 2043-2058.

CAMPBELL, N.A. AND ATCHLEY, W.R. (1981) The geometry of canonical variate analysis. Syst. Zool., 30: 268-280.

CAMPBELL, N.A. AND MAHON, R.J. (1974) A multivariate study of variation in two species of rock crab of the genus *Leptograpsus*. Aust. J. Zool., 22: 417-425.

CAMPBELL, N.A. AND REYMENT, R.A. (1978) Discriminant analysis of a cretaceous foraminifer using shrunken estimators. Math. Geol., 10: 347-359.

CANTRILL, D.J. AND WEBB, J.A. (1987) A reappraisal of *Phyllopteroides Medwell (Osmundaceae)* and its stratigraphic significance in the lower Cretaceous of eastern Australia. Alcheringa, 11: 59-85.

CHAYES, F. (1949) On ratio correlation in petrography. J. Geol., 57: 239-254.

COOLEY, W.W. AND LOHNES, P.R. (1971) Multivariate data analysis. Wiley, New York. 364p.

COLLINS, S.L.; RISSER, P.G. AND RICE, E.L. (1981) Ordination and classification of mature bottomland forests in North Central Oklahoma. Bull. Torr. Bot. Club, 108: 152-165.

CRAMER, H. (1946) Mathematical methods of statistics. Princeton University Press, Princeton. 575p.

CROVELLO, T.J. (1968) The effect of alteration of technique at two stages in a numerical taxonomic study. Univ. Kansas Sci. Bull., 47: 761-786.

DALLWITZ, M.J. (1974) A flexible computer program for generating identification keys. Syst. Zool., 33: 50-57.

DEL MORAL, R. (1975) Vegetation clustering by means of isodata: revision by multiple discriminant analysis. Vegetatio, 29: 179-190.

DIXON, W.J (Ed.) (1987) BMDP. Biomedical Computer Programs. University of California Press, Berkeley. 725p.

DODSON, P. (1978) On the use of ratios in growth studies. Syst. Zool., 27: 62-67.

DUPRAW, E.J. (1965) Non-linnean taxonomy and the systematics of honey-bees. Syst. Zool.,14: 1-24.

EASTABROOK, G.F. AND GATES, B. (1984) Character analysis in the *Banisteriopsis Campestris* complex (Malpighiaceae) using spatial auto-correlation. Taxon, 33: 13-25.

EVERITT, B.S. (1974) Cluster analysis. Heinmann, London. 122p.

EVERITT, B.S. (1979) Unresolved problems in cluster analysis. Biometrics, 35: 169-181.

FISHER, R.A. (1936) The use of multiple measurements in taxonomic problems. Ann. Eugenics, 7: 179-188.

GATTY, R. (1966) Multivariate analysis for marketing research an evaluation. Applied Statistics, 15: 157-172.

GOWER, J.C. (1966) A q-technique for the calculation of canonical variates. Biometrika, 53: 588-589.

GOWER, J.C. (1971) A general coefficient of similarity and some of its properties. Biometrics, 27: 857-872.

GOWER, J.C. (1988) Classification geometry and data analysis. In: Classification and related methods of data analysis (Ed. BOCK, H.H.), Elsevier Amsterdam. 749p.

GREEN, R.H. (1971) A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of Central Canada. Ecology, 52: 543-556.

GREEN, R.H. (1974) Multivariate niche analysis with temporally varying environmental factors. Ecology, 55: 73-83.

HARTIGAN, J.A. (1967) Representation of similarity matrices by trees. J. Amer. Stat. Assoc., 62: 1140-1158.

HAWKINS, D.M. (1974) The detection of errors in multivariate data using principal components. J. Amer. Stat. Assoc., 69: 340-344.

HEALY, M.J.R. (1968a) Algorithm AS 6. Triangular decomposition of a symmetric matrix. Applied Statistics, 17: 195-197.

HEALY, M.J.R. (1968b) Algorithm AS 7. Inversion of a positive semi-definite symmetric matrix. Applied Statistics, 17 : 198-199.

HEALY, M.J.R. (1968c) Multivariate normal plotting. Applied Statistics, 17: 157-161.

HILL, I.D. AND PIKE, M.C. (1967) Algorithm 299. Chi-squared integral. Comm. A.C.M., 10: 243.

HILL, R.S. (1980a) A numerical taxonomic approach to the study of angiosperm leaves. Bot. Gaz., 141: 213-229.

HILL, R.S. (1980b) A stopping rule for partitioning dendrograms. Bot. Gaz., 141: 321-324.

HILLS, M. (1978) On ratios- A response to Atchley, Gaskins and Anderson. Syst. Zool., 27: 61-62.

HOPPER, S.D. AND CAMPBELL, N.A. (1977) A multivariate morphometric study of species relationships in kangaroo paws (Anigozanthos Labill. and Macropidia Drumm. ex Harv. : Haemodoraceae). Aust. J. Bot., 25: 523-544.

HUMPHERIES, J.M.; BOOKSTEIN, F.L.; CHERNOFF, B.; SMITH, G.R.; ELDER, R.L. AND POSS, S.G. (1981) Multivariate discrimination by shape in relation to size. Syst. Zool., 30: 291-308.

IBBETSON, D. (1963) Algorithm 209. Gauss. Comm. A.C.M., 6: 616.

KRZANOWSKI, W.J. (1975) Discrimination and classification using both binary and continuous variables. J. Amer. Stats. Assoc., 70: 782-790.

KRZANOWSKI, W.J. (1977) The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics, 19: 191-200.

KRZANOWSKI, W.J. (1980) Mixtures of continuous and categorical variables in discriminant analysis. Biometrics, 36: 493-499.

KSHIRSAGAR, A.M. AND ARSEVEN, E. (1975) A note on the equivalency of two discrimination procedures. Amer. Statistician, 29: 38-39.

LACHENBRUCH, P. (1975) Discriminant analysis. Hafner Press, New York. 128p.

MAHALANOBIS, P.C. (1936) On the generalized distance in statistics. Proc. Nat. Inst. Sci. India, 2: 49-55.

MARDIA, K.V. (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika, 57: 519-530.

MARDIA, K.V. (1974) Applications of some measures of multivariate skewness and kurtosis to testing normality and robustness studies. Sankhya B, 36: 115-128.

MARDIA, K.V. (1975) Assessment of multivariate normality and the robustness of Hotellings T^2 test. Applied Statistics, 24: 163-171.

MARDIA, K.V. AND ZEMROCH, P.J. (1975) Measures of multivariate skewness and kurtosis. Applied Statistics, 24: 262-265.

MARKS, S. AND DUNN, O.J. (1974) Discriminant functions when covariance matrices are unequal. J. Amer. Stat. Assoc. 69: 555-559.

MAXWELL, A.E. (1961) Canonical variate analysis when the variables are dichotomous. Educ. and Psych. meas., 21: 259-271.

MAYR, E. (1969) Principles of systematic zoology. McGraw-Hill, New York. 428p.

MELTON, R.S. (1963) Some remarks on failure to meet assumptions in discriminant analysis. Psychometrika, 28: 49-51.

MORRIS, J. (1969) Algorithm 346. F-test probabilities. Comm. A.C.M., 12: 184.

NATHANSON, J.A. (1971) An application of multivariate analysis in astronomy. Applied Statistics, 20: 239-249.

NIELSON, J.S.; BROOKS, R.R.; BOSWELL, C.R. AND MARSHALL, N.J. (1973) Statistical evaluation of geobotanical and biogeochemical data by discriminant analysis. J. Appl. Ecol., 10: 251-258.

PANKHURST, R.J. (1971) Botanical keys generated by computer. Watsonia, 8: 357-368.

PAYNE, R.W. (1975) Genkey: a program for constructing diagnostic keys. In: Biological identification with computers (Ed. PANKHURST, R.J.), Academic Press, London. 333p.

PHILLIPS, B.F.; CAMPBELL, N.A. AND WILSON, B.R. (1973) A multivariate study of geographic variation in the whelk *Dicathais* J. Exp. Mar. Biol. Ecol., 11: 27-69.

PHILLIPS, R.P. (1983) Shape characters in numerical taxonomy and problems with ratios. Taxon, 32: 535-544.

PIMENTAL, R.A. (1981) A comparative study of data and ordination techniques based on a hybrid swarm of sand verbenas (Abronia juss.). Syst. Zool., 30: 250-267.

POREBSKI, O.R. (1966) Discriminatory and canonical analysis of technical college data. Brit. J. Math. Stat. Psyc., 19: 215-236.

RAO, C.R. (1947) On the significance of the additional information obtained by the inclusion of some extra variables in the discrimination of populations. Current Science, 16: 216-217.

RAO, C.R. (1950) A note on the distribution of D^2_{p+q}-D^2_p and some computational aspects of D^2 statistic and discriminant function. Sankhya, 10: 257-268.

RAO, C.R. (1952) Advanced statistical methods in biometric research. Wiley, New York. 390p.

RATKOWSKY, D.A. (1984) A stopping rule and clustering method of wide applicability. Bot. Gaz., 145: 518-523.

RATKOWSKY, D.A. AND LANCE, G.N. (1978) A criterion for determing the number of groups in a classification. Aust. Comp. J., 10: 115-117.

REYMENT, R.A. (1971) Multivariate normality in morphometric analysis. Math. Geol., 3: 357-368.

ROHLF, F.J. (1967) Correlated characters in numerical taxonomy. Syst. Zool., 16: 109-126.

SAHA, A.K. AND RAO, S.V.L.N. (1971) Quantitative discrimination between magmatic units of Singhbhum granite. Math. Geology, 3: 123-133.

SANATHANAN, L. (1975) Discriminant analysis. In: Introductory multivariate analysis, (Ed. AMICK, D.J. AND WALBERG, H.J), McCutchen. 301p.

SANGHVI, L.D. (1953) Comparison of genetical and morphological methods for a study of biological differences. Amer. J. Phys. Anthrop., 11: 385-404.

SEIDEL, M.E. AND LUCCHINO, P.V. (1981) Allozymic and morphological variation among the musk turtles *Sternotherus carnatus, S.depressus* and *S.minor* (*Kinosternidae*). Copeia, 1: 119-128.

SIMPSON, G.G. (1961) Principles of animal taxonomy. Columbia University Press, New York. 247p.

SNEATH, P.H.A. AND SOKAL, R.R. (1973) Numerical Taxonomy. Freeman, San Francisco. 573p.

SOKAL, R.R. (1965) Statistical methods in systematics. Biol. Rev., 40: 337-391.

SOKAL, R.R. AND ROHLF, F.J. (1962) The comparison of dendrograms by objective methods. Taxon, 11: 33-40.

SOKAL, R.R. AND SNEATH, P.H.A. (1963) Principles of numerical taxonomy. Freeman, San Francisco. 359p.

SOMERS, K.M. (1986) Multivariate allometry and removal of size with principal components analysis. Syst. Zool., 35: 359-368.

TITTERINGTON, D.M.; MURRAY, G.D.; MURRAY, L.S.; SPIEGELHALTER, D.J.; SKENE, A.M.; HABBEMA, J.D.F. AND GELPKE, G.J. (1981) Comparison of discrimination techniques applied to a complex data set of head injured patients. J. Roy. Stat. Soc. A, 144: 145-175.

TRUETT, J.; CORNFIELD, J. AND KANNEL, W. (1967) A multivariate analysis of the risk of coronary heart disease in Framingham. J. Chronic Disease, 20: 511-524.

WARD, J.M. (1981) Numerical phenetics and the classification of *RAOULIA* (GNAPHALIINAE-COMPOSITAE). Christchurch, University of Canterbury. 240p. (Thesis: Ph.D.: Botany)

WARD, J.M. (1982) A key, synopsis and concordance for *Raoulia*. Mauri Ora, 10: 11-19.

WEST, J.G. AND NOBLE, I.R. (1984) Analysis of digitised leaf images of the *Dodonea viscosa* complex in Australia. Taxon, 33: 595-613.

WILLIAMS, W.T. (1976) Other ordination procedures. In: Pattern analysis in agricultural science. (Ed. WILLIAMS, W.T.), Elsevier, Amsterdam. 264p.

APPENDIX A

SPECIES NAME	THESIS NAME
R.parkii Buchanan	*R.parkii*
R.australis Hooker f.	*R.australis*
R.hookeri Allan var. hookeri	*R.hookeri h.*
R.hookeri var. *albo-sericea* (Colenso) *Allan*	*R.hookeri as.*
R.hookeri var. apice nigra (Kirk) *Allan*	*R.hookeri an.*
R. beauverdii Cockayne	*R.beauverdii*
R.glabra Hooker f.	*R.glabra*
R.monroi Hooker f.	*R.monroi*
R.'sp.M' (undescribed)	*R.'sp.M'*
R.haastii Hooker f.	*R.haastii*
R.tenuicaulis Hooker f.	*R.tenuicaulis*
R.'sp.K' (undescribed)	*R.'sp.K'*
R.cinerea Petrie	*R.cinerea*
R.subsericea Hooker f.	*R.subsericea*
R.tenuicaulis var. *pusilla* Kirk	*R.tenuicaulis var. pusilla*

For a more detailed explanation of the species see Ward (1982)

APPENDIX B

OTU NO.	COLLECTION NO.	SPECIES	DIPLOID / RIVERBED
1	67501	*R.parkii*	N/N
2	67951	*R.parkii*	N/N
3	67982	*R.parkii*	N/N
4	67957	*R.parkii*	N/N
5	660101A	*R.parkii*	N/N
6	74064	*R.parkii*	N/N
7	74057-7	*R.parkii*	N/N
8	74057-4	*R.parkii*	N/N
9	67980	*R.australis*	Y/Y
10	64046	*R.australis*	Y/Y
11	64002	*R.australis*	Y/N
12	66051	*R.australis*	Y/Y
13	671146	*R.australis*	Y/Y
14	67638	*R.australis*	N/N
15	74039-6	*R.australisXparkii*	N/N
16	66100	*R.australisXhookeri h.*	N/Y
17	74022	*R.australisXhookeri h.*	N/N
18	2212	*R.hookeri h.*	N/Y
19	2206	*R.hookeri h.*	N/Y
20	74019	*R.hookeri h.*	N/Y
21	67394	*R.hookeri h.*	N/Y
22	67400	*R.hookeri h.*	N/N
23	66139	*R.hookeri h.*	N/Y
24	67942	*R.hookeri h.*	N/Y
25	67940	*R.hookeri h.*	N/Y
26	67247	*R.hookeri h.*	N/N
27	65183	*R.hookeri h.*	N/Y
28	67882	*R.hookeri h.*	N/Y
29	67941	*R.hookeri h.*	N/Y
30	66505	*R.hookeri h.*	N/N
31	671183	*R.hookeri h.*	N/N
32	74092-3	*R.hookeri h.*	N/N
33	67646	*R.hookeri as.*	N/N
34	74100	*R.hookeri as.*	N/Y
35	67752	*R.hookeri as.*	N/Y
36	67647	*R.hookeri as.*	N/N
37	67841	*R.hookeri an.*	N/N
38	68125	*R.beauverdii*	N/N
39	67037	*R.beauverdii*	N/N
40	68187	*R.hookeri an.*	N/N
41	67959	*R.hookeri an.*	N/N
42	66029B	*R.hookeri an.*	N/N
43	66594	*R.hookeri an.*	N/N
44	65284	*R.hookeri an.*	N/N
45	67954	*R.hookeri h.*	N/Y
46	68111	*R.hookeri h.*	N/Y
47	68112	*R.hookeri h.*	N/Y
48	66021	*R.hookeri h.*	N/Y
49	74024-1	*R.glabra*	Y/N
50	68045	*R.glabra*	Y/N
51	67299	*R.glabra*	Y/Y
52	65260	*R.glabra*	Y/Y

53	67981	*R.glabra*	Y/Y
54	65245A	*R.monroi*	Y/N
55	67047	*R.monroi*	Y/N
56	67470	*R.monroi*	Y/N
57	74026	*R.'sp.M'*	N/N
58	66165	*R.'sp.M'*	N/N
59	74010	*R.'sp.M'*	N/N
60	66447	*R.haastii*	Y/Y
61	66449	*R.haastii*	Y/Y
62	66475	*R.haastii*	Y/Y
63	74040-10	*R.tenuicaulis*	Y/Y
64	66476	*R.tenuicaulis*	Y/Y
65	74023	*R.tenuicaulis*	Y/Y
66	66442	*R.tenuicaulis*	Y/Y
67	65182	*R.tenuicaulis*	Y/Y
68	66529	*R.tenuicaulis*	Y/Y
69	74079-2	*R.tenuicaulis*	Y/Y
70	66461	*R.tenuicaulis*	Y/Y
71	74031-2	*R.tenuicaulis*	Y/Y
72	67607	*R.tenuicaulis*	Y/Y
73	74025-5	*R.tenuicaulis*	Y/Y
74	75010-2	*R.'sp.K'*	Y/N
75	74088	*R.'sp.K'*	Y/N
76	74089	*R.'sp.K'* × *tenuicaulis*	Y/N
77	2372LC	*R.cinerea*	-/N
78	74014	*R.subsericea*	N/N
79	67936B	*R.subsericea*	N/N
80	66010B	*R.subsericea*	N/N
81	67974	*R.subsericea*	N/N
82	68157	*R.subsericea*	N/N
83	68052	*R.subsericea*	N/N
84	68095	*R.subsericea*	N/N
85	74101	*R.australis*	N/N
86	A10188	*R.tenuicaulis var. pusilla*	Y/Y

APPENDIX C

CHARACTER LIST

VEGETATIVE:

CONTINUOUS:

1)length lamina

2)breadth lamina (at widest pt.)

3)length leaf

4)length leaf base

5)breadth leaf base (top)

6)breadth leaf base (bottom)

7)length to max. breadth from apex

8)upright shoot diameter

RATIOS:

9)lamina length/leaf length

10)lamina breadth/leaf length

11)length to max. breadth from apex/leaf length

12)lamina breadth/lamina length

13)length to max. breadth from apex/lamina length

DISCRETE:

14)shape leaf apex
- (i)truncate
- (ii)truncate to rounded
- (iii)rounded
- (iv)rounded to subacute
- (v)subacute
- (vi)subacute to acute
- (vii)acute

15)highest vein order
(i)4-5
(ii)3
(iii)2
(iv)1

16)main vein as percent lamina breadth
(i)>4%
(ii)2-4%

17)shape main vein
(i)curved
(ii)straight

18)number of traces
(i)1 strong
(ii)1 strong +1 or 2 weak laterals
(iii)1 strong +2 moderate laterals
(iv)3 strong

19)lamina folding
(i)'boated' for entire length
(ii)'boated' at tip only
(iii)none

20)tomentum coverage upper side
(i)total
(ii)margins glabrous
(iii)base glabrous
(iv)glabrous except at centre
(v)all glabrous

21)tomentum coverage lower side
(i)total
(ii)margins glabrous
(iii)base glabrous
(iv)glabrous except at centre
(v)all glabrous

22)lamina tomentum density upper side
(i)dense
(ii)moderate
(iii)thin
(iv)absent

23)lamina tomentum density lower side
(i)dense
(ii)moderate
(iii)thin
(iv)absent

24)leaf spacing on upright shoots
(i)close
(ii)intermediate
(iii)distant

25)angle of leaf to stem
(i)<45 degrees
(ii)45-90 degrees
(iii)>90 degrees

26)comparative breadths of lamina and top of leaf base
(i)lamina >0.1mm wider
(ii)lamina <0.1mm wider
(iii)variable between (ii) and (iv)
(iv)lamina <0.1mm narrower
(v)lamina >0.1mm narrower

27)tapering of leaf base
(i)top > 0.1mm wider than bottom
(ii)difference < 0.1mm
(iii)top >0.1 narrower than bottom

28)leaf base shape
(i)widest at centre
(ii)widest point non-central

29)leaf mucro form
(i)straight
(ii)upturned
(iii)absent

30)phyllotaxy
(i)spiral
(ii)two ranks

31)horizontal and upright shoots
(i)distinct
(ii)no distinction

FLORAL:

CONTINUOUS:

32)tubular floret length

33)filiform floret length

34)tubular achene length

35)filiform achene length

36)tubular achene breadth

37)filiform achene breadth

38)tubular corolla length

39)filiform corolla length

40)tubular corolla tube breadth

41)filiform corolla tube breadth

42)tubular corolla breadth at apex

43)filiform corolla breadth at apex

44)tubular corolla tube length

45)tubular corolla lobe length

46)tubular pappus length

47)filiform pappus length

48)capitulum length

49)capitulum breadth at top

50)phyllary length

51)phyllary breadth

52)length to point of max. breadth phyllary

53)number of phllaries

54)number of inner phyllaries

55)mean number of florets

56)minimum number of florets

57)maximum number of florets

58)percent filiform florets

59)tubular disc length

60)receptacle diameter

RATIOS:

61)tubular corolla tube breadth/tubular corolla length

62)filiform corolla tube breadth/filiform corolla length

63)tubular length lobe/tubular corolla length

64)tubular disc length/tubular corolla length

65)tubular corolla length/filiform corolla length

66)tubular length to point of max.breadth corolla/tubular corolla length

67)tubular pappus length-tubular corolla length

68)filiform pappus length-filiform corolla length

69)tubular corolla tube breadth/tubular corolla breadth at apex

70)filiform corolla tube breadth/filiform corolla breadth at apex

71)filiform corolla tube breadth/tubular corolla tube breadth

72)tubular achene length/filiform achene length

73)tubular achene breadth/tubular achene length

74)filiform achene breadth/filiform achene length

75)tubular floret length/filiform floret length

76)tubular achene length/tubular floret length

77)filiform achene length/filiform floret length

78)length to point of max. phyllary breadth/phyllary length

79)phyllary breadth/phyllary length

80)tubular floret length/phyllary length

81)capitulum breadth at top/capitulum length

DISCRETE:

82)shape phyllary apex
(i)truncate
(ii)truncate to rounded
(iii)rounded
(iv)rounded to subacute
(v)subacute
(vi)subacute to acute
(vii)acute

83)hairs on outer surface of inner phyllary
(i)present
(ii)absent

84)inner phyllary colour
(i)white
(ii)white to cream
(iii)cream
(iv)lemon
(v)lemon to yellow
(vi)yellow
(vii)straminous to fawn
(viii)fulvous to gold to tan

(ix)mid brown
(x)dark brown

85)receptacle curvature
(i)concave to flat
(ii)slightly rounded
(iii)rounded
(iv)conical

86)sex of tubular florets
(i)male
(ii)hermaphrodite

87)achene surface
(i)glabrous
(ii)few hairs
(iii)moderately hairy
(iv)very hairy

88)expansion of pappus hair tip
(i)great
(ii)absent
(iii)slight to moderate

89)pappus length relative to corolla length
(i)shorter
(ii)longer

90)pappus hair number
(i)14-33
(ii)approx. 50
(iii) > 80

91)pappus series number
(i)single series
(ii)several series

92)shedding of pappus hairs
(i)singly
(ii)in coherent groups

93)texture of basal part of pappus hair
(i)distinct from upper part
(ii)not distinct from upper part

94)width of body of pappus hair
(i)0.2-0.3mm
(ii)0.4-0.8mm

95)pappus hair tapering to top
(i)yes
(ii)no

96)long narrow pappillae borne at a wide angle
(i)yes
(ii)no

97)apex of pappus apical cell
(i)rounded
(ii)rounded to subacute

(iii)subacute
(iv)acute

98)corolla expansion
(i)sudden
(ii)gradual

For a more detailed explanation of characters and character states see Ward (1981).

The data matrix is in the possession of the author.

APPENDIX D

COMPUTER PROGRAMS

```
C       PROGRAM TO CALCULATE THE MULTIVARIATE MEASURES OF
C        SKEWNESS AND KURTOSIS (CM. FRAMPTON AUTHOR)
C       USING THE ALGORITHMS OF HEALY(1968A,1968B) AND MARDIA AND
C        ZEMROCH (1975)
C       RAW DATA IS CONTAINED IN MATRIX X  :N IS THE NUMBER OF OTUS
C       IP IS THE NUMBER OF VARIABLES
C       THE MEAN, VARIANCE, SKEWNESS, AND KURTOSIS FOR EACH VARIABLE
C       IS PRINTED
C       THE MAHALANOBIS D^2 OF EACH OTU TO THE CENTROID IS PRINTED
C       THE RANK OF THE INVERTED SS AND SP MATRIX IS PRINTED
C       THE FINAL MULTIVARIATE MEASURES OF SKEWNESS AND KURTOSIS
C       AND THEIR Z-SCORES ARE GIVEN
        DOUBLE PRECISIONX(63,86),S0(2020),W(63),S1(2020),XX(90,86),A(60)
        DIMENSION LL(65)
        N=86
        IP=63
        DOUBLE PRECISION XKI,XKJ,SSQM,SS,ZERO,TWO,SUMWTS,XM,B1,
        1 B2,QQ,QR,M2,M3,M4,BP1,BP2
        DATA ZERO,TWO /0D0, 2D0/
        IFAULT=0
        IF(IP .LE. 1)GOTO 860
        SUMWTS=N
        OPEN (UNIT=3,FILE='OUT.OUT',TYPE='NEW')
        DO 32 J=1,IP
        XM=ZERO
        M1=ZERO
        M2=ZERO
        M3=ZERO
        M4=ZERO
        DO 25 I=1,N
   25   XM=XM+X(J,I)
        XM=XM/SUMWTS
        DO 30 I=1,N
        X(J,I)=X(J,I)-XM
        M2=X(J,I)**2+M2
        M3=X(J,I)**3+M3
   30   M4=X(J,I)**4+M4
        M2=M2/N
        M3=M3/N
        M4=M4/N
        BP1=(M3/M2/M2**.5)**2
        BP2= M4/M2**2
   81   WRITE(3,34)J,BP1,BP2,XM,M2
   34   FORMAT(' VAR ',I2,' SKEW= ',F6.2,' KURT= ',F6.2,'MEAN=
        1 ',F6.2,'VAR= ',F8.5)
   32   CONTINUE
        L=0
        DO 40 J=1,IP
        DO 40 K=1,J
        L=L+1
        XM=ZERO
        DO 35 I=1,N
   35   XM=XM+X(J,I)*X(K,I)
        S1(L)=XM
```

```
      S0(L)=XM
40    CONTINUE
      CALL SYMINV(S1,IP,S0,W,IRANK,IFAULT,NN)
      IF(IFAULT .NE. 0)GOTO 860
      IF(IRANK .NE. 0)IFAULT=1
      IRANK=IP-IRANK
      B1=ZERO
      B2=ZERO
      DO 100 I=1,N
      XM=S0(1)*X(1,I)**2
      M=1
      DO 70 K=2,IP
      QQ=ZERO
      KK=K-1
      DO 50 L=1,KK
      M=M+1
      QQ=QQ+S0(M)*X(L,I)
50    CONTINUE
      M=M+1
      XKI=X(K,I)
      XM=XM+(S0(M)*XKI+TWO*QQ)*XKI
70    CONTINUE
      QQ=XM**2
      B2=B2+QQ
      B1=B1+QQ*XM
      WRITE(3,955)I,XM*N
955   FORMAT(' D2 BET OTU ',I2,' & MEAN = ',F6.2)
100   CONTINUE
      DO 150 I=2,N
      II=I-1
      SS=S0(1)*X(1,I)
      DO 140 J=1,II
      XM=SS*X(1,J)
      M=1
      DO 120 K=2,IP
      XKI=X(K,I)
      XKJ=X(K,J)
      QQ=ZERO
      QR=ZERO
      KK=K-1
      DO 110 L=1,KK
      M=M+1
      SSQM=S0(M)
      QQ=QQ+SSQM*X(L,J)
      QR=QR+SSQM*X(L,I)
110   CONTINUE
      QQ=QQ*XKI+QR*XKJ
      M=M+1
      XM=XM+QQ+S0(M)*XKI*XKJ
120   CONTINUE
      B1=B1+TWO*XM**3
140   CONTINUE
150   CONTINUE
      B1=B1*SUMWTS
      B2=B2*SUMWTS
      WRITE(3,950)B1,B2,IRANK
950   FORMAT(' B1= ',F9.2,' B2= ',F9.2,' RANK:P= ',I3)
      AA=N*B1/6
      DF1=IRANK*(IRANK+1)*(IRANK+2)
      BN1=SQRT(2*AA)-SQRT((DF1-3)/3)
      DF2=IRANK*(IRANK+2)
```

```
      BN2=(B2-DF2*(N-1)/(N+1))/SQRT(8*DF2/N)
      WRITE(3,111)BN1,BN2
111   FORMAT(' Z-SCORE B1= ',F11.2,' Z-SCORE B2= ',F8.2)
750   CLOSE (UNIT=3)
      STOP
860   WRITE(3,850)IFAULT
850   FORMAT(' ERROR=',I2)
      GOTO 750
      END
      SUBROUTINE SYMINV(S1,IP,S0,W,IRANK,IFAULT,NN)
      DOUBLE PRECISION S0(2020),S1(2020),W(2020)
      NROW=IP
      IFAULT=1
      IF(NROW .LE. 0)GOTO 100
      IFAULT=0
      CALL CHOL(S1,IP,S0,IRANK,IFAULT)
      IF(IFAULT .NE. 0)GOTO 100
      IROW=NROW
      NDIAG=NN
16    IF(S0(NDIAG) .EQ. 0.0)GOTO 11
      L=NDIAG
      DO 10 I=IROW,NROW
      W(I)=S0(L)
      L=L+I
10    CONTINUE
      ICOL=NROW
      JCOL=NN
      MDIAG=NN
15    L=JCOL
      X=0.0
      IF(ICOL .EQ. IROW) X=1.0/W(IROW)
      K=NROW
13    IF(K .EQ. IROW)GOTO 12
      X=X-W(K)*S0(L)
      K=K-1
      L=L-1
      IF(L .GT. MDIAG)L=L-K+1
      GOTO 13
12    S0(L)=X/W(IROW)
      IF(ICOL .EQ. IROW)GOTO 14
      MDIAG=MDIAG-ICOL
      ICOL=ICOL-1
      JCOL=JCOL-1
      GOTO 15
11    L=NDIAG
      DO 17 J=IROW,NROW
      S0(L)=0.0
      L=L+J
17    CONTINUE
14    NDIAG=NDIAG-IROW
      IROW=IROW-1
      IF(IROW .NE. 0)GOTO 16
100   RETURN
      END
      SUBROUTINE CHOL(S1,IP,S0,IRANK,IFAULT)
      DOUBLE PRECISION S0(2020),S1(2020),ETA
      DATA ETA /1.0D-5/
      IFAULT=1
      IF(IP .LE. 0)GOTO 100
      IFAULT=2
      IRANK=0
```

```
      J=1
      K=0
      DO 10 ICOL=1,IP
      L=0
      DO 11 IROW=1,ICOL
      K=K+1
      W=S1(K)
      M=J
      DO 12 I=1,IROW
      L+1
      IF(I .EQ. IROW)GOTO 13
      W=W-S0(L)*S0(M)
      M=M+1
12    CONTINUE
13    IF(IROW .EQ. ICOL)GOTO 14
      IF(S0(L) .EQ. 0.0)GOTO 21
      S0(K)=W/S0(L)
      GOTO 11
21    S0(K)=0.0
11    CONTINUE
14    ZX=ETA*S1(K)
      IF(ABS(W) .LT. ABS(ZX))GOTO 20
      IF(W .LT. 0.0)GOTO 100
      S0(K)=SQRT(W)
      GOTO 15
20    S0(K)=0.0
      IRANK=IRANK+1
15    J=J+ICOL
10    CONTINUE
      IFAULT=0
100   RETURN
      END
```

```
C       C.A.G.S  A NEW MEASURE FOR THE COMPARATIVE ASSESSMENT OF
C       GROUPING STRATEGIES IN A MULTIVARIATE DATA SET OF MIXED
C       VARIABLE TYPES. (AUTHOR CM FRAMPTON)
C       GNG=THE NUMBER OF GROUPS
C       NC=THE NUMBER OF OTUS
C       NV=THE NUMBER OF CHARACTERS
C       W CONTAINS THE RAW DATA
C       THE GROUPING VARIABLE IS W(G,*)
C       THE INDIVIDUAL VARIABLE CONTRIBUTIONS TO THE FINAL STATISTIC
C       ARE W(*,100)
C       X CONTAINS THE MINIMUM AND MAXIMUM STATES FOR EACH
C       DISCRETE VARIABLE X(*,1) AND X(*,2) RESPECTIVELY
C       X(*,0) IS 0 FOR CONTINUOUS CHARACTER AND 1 FOR A DISCRETE
C       CHARACTER
        DIMENSION W(100,100),X(100,2),RT(20),CT(20),Z(12,20),GT(20),GN(20),
        GNG=12
        NC=85
        TX2=0
        TDF=0
        DO 1000 J,NV
        IF (X(J,0) .EQ. 1)GOTO 200
C       FOR CONTINUOUS CHARACTERS
        BSS=0
        TT=0
        TS=0
        DO 54 M=1,20
        GT(M)=0
   54   GN(M)=0
        DO 50 PN=1,NC
        L=W(G,PN)
        GT(L)=GT(L)+W(J,PN)
        GN(L)=GN(L)+1.0
        TT=TT+W(J,PN)
        TS=TS+W(J,PN)**2
   50   CONTINUE
        DO 60 M=1,GNG
   60   BSS=BSS+GT(M)**2/GN(M)
        CTT=TT**2/NC
        FF=(BSS-CTT)/(TS-CTT)
        FF=1/FF-1
        FF=1/FF*(NC-GNG-2)/(GNG-1)
        W(J,100)=FF
        TDF=TDF+GNG-1
        TX2=TX2+(W(J,100)*(GNG-1))
        GOTO 1000
C       FOR DISCRETE CHARACTERS
  200   CONTINUE
        DO 67 K=1,20
        RT(K)=0.0
        CT(K)=0.0
        DO 67 R=1,13
   67   Z(R,K)=0.0
        CS=0
        DO 90 PN=1,NC
        A=W(J,PN)+1
        L=W(G,PN)
   90   Z(A,L)=Z(A,L)+1
        DO 95 O=1,GNG
        DO 98 M=X(J,1)+1,X(J,2)+1
        RT(O)=Z(M,O)+RT(O)
```

```
      CT(M)=Z(M,O)+CT(M)
98    CONTINUE
95    CONTINUE
      DO 120 M=X(J,1)+1,X(J,2)+1
      DO 130 O=1,GNG
      EF=RT(O)*CT(M)/NC
      CS=(Z(M,O)-EF)**2/EF+CS
130   CONTINUE
120   CONTINUE
      LC=X(J,2)-X(J,1)
      TDF=TDF+(LC*(GNG-1))
      TX2=TX2+CS
      IF(LC .GT. (GNG-1))LC=GNG-1
      W(J,100)=CS
1000  CONTINUE
      ZZ=SQRT(2*TX2)-SQRT(2*TDF-1)
      OPEN (UNIT=3,FILE='OUT.OUT',TYPE='NEW')
      WRITE(3,7000),TDF,TX2,ZZ
7000  FORMAT(' TDF= ',F9.1,'TX2= ',F10.2,'Z= ',F7.2)
      DO 150 MP=1,NV
      WRITE(3,45)MP,W(MP,100)
150   CONTINUE
45    FORMAT(I2,2X,F8.2)
      CLOSE (UNIT=3)
      STOP
      END
```

```
(Pascal functions for statistical probabilities)
(Author CM Frampton: using the algorithms of Ibbetson(1963),Morris(1969), and Hill & Pike(1967)
(Gauss returns the upper tail normal probability for abs(x))
      Function gauss(x : Double) : Double;
      Var y,z,w : double;
      Begin
            If x=0.0 Then z := 0.0
                  Else Begin
                        y := abs(x)/2.0;
                        If y>=3 Then z := 1.0
                        Else If y<1.0 Then
                              Begin
                                    w := y*y;
                                    z := (((((((0.000124818987*w
                                          -0.001075204047)*w+0.005198775019)*w
                                          -0.019198292004)*w+0.059054035642)*w
                                          -0.151968751364)*w+0.319152932694)*w
                                          -0.531923007300)*w+0.797884560593)*y
                                          *2.0;
                              End Else
                              Begin
                                    y := y-2.0;
                                    z := ((((((((((((-0.000045255659*y
                                          +0.000152529290)*y-0.000019538132)*y
                                          -0.000676904986)*y+0.001390604284)*y
                                          -0.000794620820)*y-0.002034254874)*y
                                          +0.006549791214)*y-0.010557625006)*y
                                          +0.011630447319)*y-0.009279453341)*y
                                          +0.005353579108)*y-0.002141268741)*y
                                          +0.000535310849)*y+0.999936657524
                              End
                  End;
If x>0.0 Then gauss := (z+1.0)/2.0
            Else gauss := (1.0-z)/2.0;
End;

(Chiprob returns the upper tail probability of the chi-square distribution for <chi> given degrees
(of freedom df)
Function chiprob(chi : Double;df : Integer) : Double;
Var a,y,s,e,c,z : Double;
      even,bigchi : Boolean;
Begin
y := 0; { dummy value }
bigchi := chi>205;
a := 0.5*chi;
even := Not Odd(df);
If even Or (df>2) And (not bigchi) Then y := exp(-a);
If even Then s := y
      Else s := 2.0*gauss(-sqrt(chi));
If df>2 Then
      Begin
            chi := 0.5*(df-1.0);
            If even Then z := 1.0
                  Else z := 0.5;
            If bigchi Then
                  Begin
                        If even Then e := 0
                              Else e := 0.572364942925;
                        c := ln(a);
                        While z<=chi Do
```

```
                    Begin
                        e := ln(z)+e;
                        s := exp(c*z-a-e)+s;
                        z := z+1.0;
                    End;
                chiprob := 1.0-s;
            End
            Else Begin
                If even Then e := 1.0
                    Else e := 0.564189583548/sqrt(a);
                c := 0;
                While z<=chi Do
                    Begin
                        e := e*a/z;
                        c := c+e;
                        z := z+1.0;
                    End;
                chiprob := 1.0-(c*y+s);
            End
        End Else chiprob := 1.0-s;
    End;

(Fprob returns the upper tail probability of the F distribution for <F> with degrees of freedom
(<df1> and <df2>)
Function fprob(f : Double;df1,df2 : Integer) : Double;
Var maxn : Integer;
    f1,f2,x,xx,ft,vp : Double;
    theta,sth,cth,sts,cts,a,b,xi,gamma : Double;
    cbrf : Double;
Begin
maxn := 500;
If f>100000 Then fprob := 0.0
    Else
If f=0.0 Then fprob := 1.0
    Else
If df1=1 Then fprob := tprob(sqrt(f),df2)
    Else
If df2=1 Then fprob := 1.0-tprob(sqrt(1/f),df1)
    Else Begin
        f1 := df1;
        f2 := df2;
        ft := 0.0;
        x := f2/(f2+f1*f);
        vp := f1+f2-2.0;
        If (2*(df1 div 2)=df1) And (df1<=maxn) Then
            Begin
                xx := 1.0-x;
                f1 := f1-2.0;
                While f1>=1.0 Do
                    Begin
                        vp := vp-2.0;
                        ft := xx*vp/f1*(1.0+ft);
                        f1 := f1-2.0;
                    End;
                ft := x**(0.5*f2)*(1.0+ft)
            End
            Else If (2*(df2 div 2)=df2) And (df2 <= maxn) Then
            Begin
                f2 := f2-2.0;
                While f2>=1.0 Do
                    Begin
```

```
                    vp := vp-2.0;
                    ft := x*vp/f2*(1.0+ft);
                    f2 := f2-2.0;
                End;
            ft := 1.0-(1.0-x)**(0.5*f1)*(1.0+ft)
        End
        Else If df1+df2<= maxn Then
        Begin
            theta := arctan(sqrt(f1*f/f2));
            sth := sin(theta);
            cth := cos(theta);
            sts := sth**2;
            cts := cth**2;
            a := 0.0;
            b := 0.0;
            If df2>1 Then
                Begin
                    f2 := f2-2.0;
                    While f2>1.0 Do
                        Begin
                            a := cts*(f2-1.0)/f2*(1.0+a);
                            f2 := f2-2.0;
                        End;
                    a := sth*cth*(1.0+a);
                End;
            a := theta+a;
            If df1>1 Then
                Begin
                    f1 := f1-2.0;
                    While f1>=2.0 Do
                        Begin
                            vp := vp-2.0;
                            b := sts*vp/f1*(1.0+b);
                            f1 := f1-2.0;
                        End;
                    gamma := 1.0;
                    f2 := 0.5*df2;
                    xi := 1.0;
                    While xi<=f2 Do
                        Begin
                            gamma := xi*gamma/(xi-0.5);
                            xi := xi+1.0;
                        End;
                    b := gamma*sth*cth**df2*(1.0+b);
                End;
            ft := 1.0+0.636619772368*(b-a);
        End
        Else Begin
            f1 := 2.0/(9.0*f1);
            f2 := 2.0/(9.0*f2);
            cbrf := f**0.333333333333;
            ft := gauss(-((1.0-f2)*cbrf+f1-1.0)/sqrt(f2*cbrf*cbrf+f1));
        End;
    If ft<0.0 Then fprob := 0.0 Else fprob := ft;
End
End;

(Cprob returns a chi square value given an upper tail probability <a> degrees of freedom <df> and
(an upper and lower estimate of cprob)
Function cprob(a,es1,es2 : Double ; df:integer): Double;
Var a1,a2,ae,cest : Double;
```

```
Begin
      a1 := 1-Chiprob(es1,df);
      a2 := 1-Chiprob(es2,df);
      cest := (es2-es1)/(a2-a1)*(a-a1)+es1;
      ae := 1-Chiprob(cest,df);
            While abs(a-ae)>1.0D-8 Do
            Begin
                  If ae<a Then es2 := cest
                  Else es1 := cest;
            cest := (es2-es1)/(a2-a1)*(a-a1)+es1;
            ae := 1-Chiprob(cest,df);
            end;
      cprob := cest
end;
```